Advanced Calculus

David Fearnley, D. Phil Oxford

fearnlda@uvu.edu

Contents

Chapter 1

The Field of Real Numbers

Foreword: The objective of this particular advanced calculus text is to present the basics of Advanced Calculus at a level which is appropriate for an undergraduate student who has taken few proof classes. The primary objective is not to cover a full grounding in the subject, but to cover a sufficient foundation to understand basic ideas without many detours, so as to keep what is covered concise and understandable. Since the single variable portion of the text is intended as more of a stand-alone text than a foundation for later analysis courses, we focus somewhat more on sequence arguments and somewhat less on more involved topological ideas. Enrichment topics are listed at the end for those who would like to incorporate more of an associated idea than the planned core topics listed. We use topological ideas more in the later part of the text.

As with most advanced calculus and analysis texts, the proofs themselves are informed by many sources. While the proofs are written by the author, no theorem or proof approach in this text is actually due to the author (they are all standard theorems with presumably typical proof approaches as found in assorted texts in the literature, most of which cannot really be tracked down to an original author who first presented a given method of argument). Though I have mentioned references to a few of the longer arguments that are particularly close to those found in other sources, I have not normally made an effort to find original sources to particular proofs for most theorems. I apologize in advance to any to whom I may not have given adequate attribution.

Primitive Notions

Before we discuss the main body of the subject matter, it is probably appropriate to mention that all of these proofs assume the rules of logic and the axioms of set theory at their foundation. Since we are intentionally attempting not to wander too far into tangents, we will just say that mathematics has some fundamental assumptions that it is okay to take unions of sets and define subsets of sets consisting of elements satisfying certain statements, and take sets of subsets of sets and choose a set of elements consisting of one element from each of a collection sets that are known to be non-empty, and that mathematical proofs should follow intuitively clear logical assumptions like ”if it is true that whenever statement A is true, statement B is true, and it is also true that whenever statement B is true, statement C is true, then it must follow that whenever statement A is true, statement C is true.” We will not formalize these ideas in this book.

Definition 1

The notation $x \in S$ indicates that x is an *element* of the *set* S. We think of a set as a collection of points. Point is often another term we use for an element of a set (particularly if that set is the real numbers or Euclidean space of any other dimension). We say that $S \subseteq T$ (meaning S is contained in T or is a *subset* of S) if every element of S is an element of T. We use $S \subset T$ to mean that S is a proper subset of T (a subset which is not equal to T). Two sets are *equal* if they have the same elements.

After the notion of a set, one of the most fundamental notions is that of a function, which we define next.

Definition 2

The notation $A \times B$ (often read "A cross B") refers to the set of ordered pairs (a, b) so that $a \in A$ and $b \in B$. This is referred to as the *Cartesian product* ot A and B or as *the cross product* of A and B.

A *function* $f : A \to B$ is a subset of $A \times B$ so that for every $a \in A$ there is exactly one $b \in B$ so that $(a, b) \in f$. Normally, instead of writing "$(a, b) \in f$" we use the notation $f(a) = b$. In this way, we may think of f as a way of assigning a point of B to each point of A. We also say that if $f(a) = b$ then b is the image of point a. If $g : B \to C$ then we define the *composition* $g \circ f : A \to C$ to be the set of all points (a, c) of $A \times C$ so that $(a, b) \in f$ and $(b, c) \in g$. This is also written $g \circ f = g(f)$, so we would say that $g \circ f(a) = c$ or $g(f(a)) = c$ if $(a, c) \in g \circ f$.

Associated with a function we can talk about the image of a set or the *inverse image* (or pre-image) of a set. If $U \subseteq A$ we define $f(U) = \{b \in B | f(a) = b$ for some $a \in U\}$. We define $f^{-1}(V) = \{a \in A | f(a) \in V\}$. Note that the definition of inverse of a set does not require a function to be invertible for sets to have inverse images. We call $f(A)$ the *range* of f (denoted $ran(f)$) and A the *domain* of f (denoted $dom(f)$). The *implied domain* of an expression for a function with real numbers as the codomain (a function for which the formula gives real valued outputs) is the set of real numbers for which the function could be defined by the expression given for the function. This is understood to be the domain when no domain is specified. The implied domain of $f+g$ and fg is $dom(f) \cap dom(g)$ for two real valued functions f, g, and the implied domain of $\frac{f}{g}$ is $\{x \in (dom(f) \cap dom(g)) | g(x) \neq 0\}$, and the implied domain of $f \circ g$ is $\{x \in dom(g) | g(x) \in dom(f)\}$.

We say that f is *one to one* if whenever $x \neq y$ and $x, y \in dom(f)$, $f(x) \neq f(y)$. This can also be stated as saying that if $f(x) = f(y)$ then $x = y$.

We say $f : A \to B$ is *onto* (with respect to specified codomain B) if for each $b \in B$ there is an $a \in A$ so that $f(a) = b$. Another way of saying this is that $ran(f) = B$. A function which is one to one and onto is a *one to one correspondence*. If $f : A \to B$ is a one to one correspondence then we define the *inverse function* of f or just the inverse of f to be the function $f^{-1} : B \to A$ defined by $f^{-1} = \{(b, a) \in B \times A | (a, b) \in f\}$, and we say that f is *invertible.*

Discussions about sets will frequently require being able to combine or separate them, so we next define union and intersection.

Definition 3

The notation $A \cup B$ means the *union* of A and B, and is the set of points which are in at least one of the sets A or B. The notation $A \cap B$ means the *intersection* of A and B, and is the set of points which are in both set A and set B. The *complement* of set B in set A is denoted $A \setminus B$ and is the set of points of A which are not points of B. If U_α is a set for every $\alpha \in J$ (where J is an unspecified set of indices) then we can use the notation $\bigcup\limits_{\alpha \in J} U_\alpha$ to denote the union of all sets U_α, which is the set of points in at least one of the U_α sets. We can define intersection over arbitrarily many indices similarly. A point $x \in \bigcap\limits_{\alpha \in J} U_\alpha$ if x is an element of every set U_α. If C is a set of sets then $\bigcup C$ refers to the union of all elements of C and $\bigcap C$ refers to the intersection of all elements of C. If $\mathbb{N}$ represents the set of natural numbers (to be defined later), then $\bigcap\limits_{n \in \mathbb{N}} A_n$ and $\bigcap\limits_{n=1}^{\infty} A_n$ mean the same thing. Likewise, $\bigcup\limits_{n \in \mathbb{N}} A_n = \bigcup\limits_{n=1}^{\infty} A_n$.

If $\mathcal{C} = \{U_\alpha\}_{\alpha \in J}$ is a collection of sets so that $U_\alpha \cap U_\beta = \emptyset$ for all $\alpha \neq \beta$ in J then we say that C is a *pairwise disjoint* collection of sets or that the U_α sets are pairwise disjoint. We may just say the collection of sets is disjoint and leave off the ”pairwise.” For instance, we may say that sets A and B are disjoint rather than saying $\{A, B\}$ is a pairwise disjoint collection of sets.

In general, if we use an unspecified set like J in the statement of a theorem without specifying what J is, then it should be understood that the statement is being said to be true for any arbitrary set J. It is worth mentioning that the notation $A \subset B$ is frequently used in mathematical texts to mean $A \subseteq B$, but we will not adopt that convention for this book.

Note that from a strictly set construction sort of point of view it is somewhat lacking in rigor to simply define an ordered pair and assume it exists. Rather, we would say an ordered pair (a, b) is the set $\{\{a, b\}, \{b\}\}$. However, it is more convenient to write an ordered pair as just (a, b). Also, in some texts the term ”mapping” may refer to a continuous function or homomorphism, but we will be using the word to refer to any function.

We have not defined the real numbers, but readers will have a pretty good idea what the real number system represents so we will use that set to illustrate the set notions discussed above despite the fact that it is not defined yet. So, for the next two examples, we will assume that we already understand what the real numbers and natural numbers are.

Example 1.1. *Let $A = [-1, 4]$ and let $B = (3, 5)$, and let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. Determine:*

(a) $A \cup B$
(b) $A \cap B$
(c) $A \setminus B$
(d) $f(A)$
(e) $f^{-1}(A)$
(f) $dom(f)$
(g) $ran(f)$
(h) Is f one to one?
(i) Is f onto?

Solution. (a) $A \cup B = [-1, 5)$
(b) $A \cap B = (3, 4]$
(c) $A \setminus B = [-1, 3]$
(d) $f(A) = [0, 16]$
(e) $f^{-1}(A) = [-2, 2]$
(e) $dom(f) = \mathbb{R}$
(f) $ran(f) = [0, \infty)$.
(h) f is not one to one because $f(x) = f(-x)$ for all $x \in \mathbb{R}$.
(i) f is not onto because f is written as $f : \mathbb{R} \to \mathbb{R}$ so that the listed codomain is not the same as the range.

□

Notice that whether a function is onto is a property of the specified codomain for the function. The function itself is the same set of ordered pairs whether we write $f : \mathbb{R} \to \mathbb{R}$ or $f : \mathbb{R} \to [0, \infty)$. However, if we use the first notation then the specified codomain is $\mathbb{R}$ which contains $[0, \infty)$ as a proper subset. Since $ran(f) = [0, \infty) \subset \mathbb{R}$, in the notation $f : \mathbb{R} \to \mathbb{R}$, the function f is not onto, but if we designate the function as $f : \mathbb{R} \to [0, \infty)$ instead then the function is onto with respect to the new codomain (even though it is exactly the same function).

Example 1.2. *Let $A_n = [-1 - \frac{1}{n}, 1 + n]$ for each $n \in \mathbb{N}$ (in other words $n = 1, 2, 3, 4, ...$). Find*

(a) $\bigcup_{n=1}^{\infty} A_n$
(b) $\bigcap_{n \in \mathbb{N}} A_n$.

Solution. (a) $\bigcup_{n=1}^{\infty} A_n = [-2, \infty)$
(b) $\bigcap_{n \in \mathbb{N}} A_n = [-1, 2]$.

□

Truth Tables, Logic and ZFC

A statement is symbolized with a letter in prepositional calculus. The symbol $\wedge$ means "and" and $\vee$ means "or." Saying "statements p and q are both true" is symbolized by $p \wedge q$ and is referred to as a conjunction. Saying "statement p is true or statement q is true (or both)" is symbolized by $p \vee q$ and is referred to as a disjunction. Statements p and q are thought of as "atoms" in these statements, or statements on whose truth the truth of the conjunction or disjunction depends. In mathematics "or" always means "inclusive or," so if either statement is true or both are true then the disjunction is true. We actually very rarely talk about the terms "conjunction" or "disjunction" except in discrete mathematics and formal logic. We are only presenting this brief introduction to give students a notion of logical inference to better understand what is meant by a proof. For some students the symbols will be useful, for others they will add confusion, but the formality provides a structure that illustrates the type of rigor we hope to see in arguments using words. Logical negation $\neg p$ being true means "p is false." We assume the following truth table without proof (it is a principle of logic we cannot prove, and is essentially how we define when "and" and "or" are true based on their atomic statements).

All "statements" in this text that are referred to are assumed to be well defined and properly formulated within the language of logic and the set theory and are either true or false but not both (we will not refer to sentences like "this statement is false" as a "statement" for example).

We will refer to this as the Primitive Truth Table:

p	q	$p \wedge q$	$p \vee q$	$\neg p$
T	T	T	T	F
T	F	F	T	F
F	T	F	T	T
F	F	F	F	T

This truth table can be used with other basic statements about logic to derive rules of inference under the assumption that any statement can be used as an atomic statement and conjunctions and disjunctions placed back into the truth table to determine its truth based on the truth of the atomic statements any number of times. We use $p \to q$ to indicate "if p is true then q is true." This is equivalent to the statement $\neg p \vee q$ because if we use a truth table then we see that whatever entries of true or false we assign to p and q, $\neg p \vee q$ is true whenever $p \to q$ is true, and vice versa. Two logical statements a and b are *equivalent* if one is true if and only if the other is true, in which case we sometimes indicate this with the symbol $a \leftrightarrow b$. To save time, since our logic development is intended to be fairly brief, we are simply going to define "if p is true then q is true" to mean $\neg p \vee q$ is a true statement, which is why we do not have an entry for $p \to q$ on our primitive truth table. If p is true and $\neg p \vee q$ is a true statement then since $\neg p$ is false it follows that q must be true. More formally, on the following truth table we see that whenever p is true it is true that q is true if and only if $\neg p \vee q$ is a true statement. This truth table is one in which we use a portion of the preceding truth table to determine the truth of $\neg p$ and then use the $\neg p$ and q columns for the p and q columns in the first truth table to obtain the entries in the $\neg p \vee q$ column.

p	q	$\neg p \vee q$	$\neg p$
T	T	T	F
T	F	F	F
F	T	T	T
F	F	T	T

A statement which is always true is called a *tautology*, such as $p \vee \neg p$. A statement which is never true is a *contradiction*, such as $p \wedge \neg p$. Here are a collection of standard rules of inference with names. You are not required to memorize the names of these rules. Verifying a rule of inference can be done by using the primitive truth table's assumptions and making columns for the statements in the hypotheses using the atoms in the hypotheses and then using these larger statements as atoms in the same truth table, showing that whenever the full statement which is the hypothesis is true, the conclusion is also true. In this manner, we determine that the implication given is a tautology (it is always true) and is thus a valid rule of inference. It is also fine to use earlier established rules of inference to derive other rules of inference without using a truth table.

Modus Ponens: $(p \wedge (p \to q)) \to q$
Modus Tollens: $(\neg q \wedge (p \to q)) \to \neg p$
Reductio Ad Absurdum: $((p \to q) \wedge (p \to \neg q)) \to \neg q$.
Noncontradiction: $(p \wedge \neg p) \to q$
Double Negation: $\neg\neg p \to p$
Case Analysis: $((p \vee q) \wedge (p \to r) \wedge (q \to r)) \to r$
Disjunctive Syllogism: $((p \vee q) \wedge (\neg p)) \to q$
Constructive Dilemma: $((p \to r) \wedge (q \to s) \wedge (p \vee q)) \to (r \vee s)$
Absorption: $(p \to q) \to (p \to (p \wedge q))$
Hypothetical Syllogism: $((p \to q) \wedge (q \to r)) \to (p \to r)$

These are just rules for propositional logic. This is entirely inadequate for arguments in mathematics, and in and of itself it models almost none of the proofs in advanced calculus. First order logic also includes the ideas of sets and quantifiers. In this discussion we are allowed to use statements that depend on a variable so that we can say that a statement that is true for some or all elements of a set, and if P is a statement about elements x then $P(x)$ means the statement P is true about x. The statement "for every point x in a set A, the statement $P(x)$ is true" is written $\forall x \in A(P(x))$ and "there exists some $x \in A$ so that $P(x)$ is true" is written $\exists x \in A(P(x))$. Similar to the preceding truth table we have as an assumption that $\neg(\forall x \in A(P(x))) \leftrightarrow (\exists x \in A(\neg P(x))$. A variable x which a statement's truth depends on is *bound* if it is in a quantifier, and *free* otherwise. For instance, in the statement "for every real number x it is true that $x^2 > y$", the variable x is bound because it is the variable for the quantifier "for every x", whereas the variable y is free. By adding these quantifiers in, we are now able to make the majority of statements that one sees in an advanced calculus book. Generally, an atomic statement is appended for each value in a given set about which a statement with a universal quantifier or statement of existence is used.

While it can be useful to use this sort of notation to keep your arguments straight at times, most of the time it is easier to understand and easier to prove statements using words, but if your proof is not clear enough that it could be replaced by symbolic logic statements in an obvious way, then there is a good chance you have not written out a complete proof. What things must be included varies depending on background and audience and the place in a sequence of arguments in which a proof occurs. For instance, at the beginning of our development when we are proving statements about field axioms, care must be taken to cite every step and justification and axiom. Later in the course we will not refer to a rule when dividing both sides of an equation by a number or cancel two objects in a division because it is assumed that this foundation is understood and well known at that point, but we should be able to reduce the argument to axioms and logical rules if asked. If we cannot then we probably have not written a rigorous argument.

One consequence of the preceding rules is that when creating a negation of a statement (a statement which is true if and only if the original statement is false) we can simply replace $\forall$ be $\exists$ and $\exists$ by $\forall$ and replace $\wedge$ by $\vee$ and $\vee$ by $\wedge$ and replace all atomic statements by their negations. For example, later in this text we will learn that if c is a limit point of the domain D of a function f, then the definition of $\lim_{x \to c} f(x) = L$ is that for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |x-c| < \delta$ and $x \in D$ then $|f(x) - L| < \epsilon$. If we were to write that as a statement in terse logical notation and we let P be the set of positive numbers, we could write $\forall \epsilon \in P(\exists \delta \in P(\forall x \in D(|f(x) - L| < \epsilon \vee \neg(0 < |x-c| < \delta))))$, with negation $\exists \epsilon \in P(\forall \delta \in P(\exists x \in D((0 < |x-c| < \delta) \wedge \neg(|f(x) - L| < \epsilon))))$. In words, that would be "for some $\epsilon > 0$, for each $\delta > 0$ there is some real number $x \in D$ so that $0 < |x-c| < \delta$ but $|f(x) - L| \geq \epsilon$." Note that "but" means the same thing as "and" in mathematics, and the main reason to use one rather than the other is to draw attention to certain implications of the statement to the reader.

The Axioms of Zermelo-Frankel Set Theory with Choice (ZFC) are (in words as intuitive descriptions, and not in any particular order):

Axiom 0: Existence. There is a set.

Axiom 1: Pairing. If A and B are sets then there is a set $\{A, B\}$.

Axiom 2: Union. If $\mathcal{C}$ is a collection of sets then there is a set containing every point which is an element of at least one element of $\mathcal{C}$.

Axiom 3: Extensionality. Two sets are equal if and only if they contain the same elements.

Axiom 4: Foundation. Every set A contains an element B which does not share any elements with A.

Axiom 5: Separation. For a well defined statement ϕ about a point (which is true or false but not both) for each element of a set S, the subset $A = \{x \in S | \phi(x) \text{ is true}\}$ exists.

Axiom 6: Infinity. There is a set ω containing the empty set as an element having the property that for every $\alpha \in \omega$, the point $\alpha \cup \{\alpha\} \in \omega$.

Axiom 7: Schema of Replacement. If S is a set and ϕ is a well defined statement so that for each $\alpha \in S$ there is exactly one β such that $\phi(\alpha, \beta)$ is true, then the set $T = \{\beta | \phi(\alpha, \beta)$

is true for some $\alpha \in S\}$ exists.

Axiom 8: Power Set. If A is a set then there is a set containing all subsets of A (and hence, by Specification, there is a set consisting of exactly the subsets of A, which we refer to as the power set of A).

Axiom 9: Choice. Every set can be linearly ordered in such a way that every non-empty subset has a least element.

This is also written as: If A is a set of non-empty sets then there is a set B consisting of exactly one point from each element of A.

We will not be referring to these axioms or to these logical rules, but in the back of our minds it is perhaps lacking in rigor to not at least mention that these concepts are being used all the time (every time we define a set or take a union for instance, we are making use of an assumption listed above). The two forms of Choice above can be shown to be equivalent, but this would detract from the main point of the text (it is a somewhat lengthy set theory argument). The second form of the axiom is the one most needed for our arguments, but the first form proves the second much more easily than the second form can be used to prove the first. In fact, if we just use the well-ordering (first) form of the axiom, take all sets in a collection S of sets and then well-order the union of S using the axiom, the set consisting of the first element of each of the elements of S (which exists by Replacement Schema) is a set of the form specified in the second form of the axiom. Thus, though it is less natural, we will use the first form of the axiom as the Axiom of Choice for our development though, as mentioned, we never actually refer to it in advanced calculus (we just use it without saying so).

Field Axioms

Definition 4

A set S together with functions $+, \cdot : S \times S \to S$ (called *binary operations* on S), is a *field* if it satisfies the following requirements, which we will refer to as axioms of a field. Note that rather than writing $+(a, b) = c$ we write $a + b = c$ (read a plus b equals c), and rather than writing $\cdot(a, b) = c$ we write $ab = c$ (read a times b equals c), and use parenthesis to indicate order of operations. Operations within parentheses are understood to occur first, and otherwise multiplication is understood to be applied before addition. Thus, $ab + c$ means $+(\cdot(a, b), c)$, for example.

For all $a, b, c \in S$ the following are true:

Commutativity: $a + b = b + a$ and $ab = ba$

Associativity: $a + (b + c) = (a + b) + c$ and $a(bc) = (ab)c$

Distributivity: $a(b + c) = ab + ac$

Identity: There are distinct elements $0, 1 \in S$ so that for any $a \in S$, $a + 0 = a$ and $(a)(1) = a$.

Inverses: For any $a \in S$ there is a point $-a \in S$ (called the additive inverse of a) so that $a + -a = 0$. If $a \neq 0$ then there is a point $\frac{1}{a} \in S$ (called the multiplicative inverse of a) so that $(a)(\frac{1}{a}) = 1$.

For some, using these things as starting assumptions should be motivated, but any attempt to do so will likely be motivation of an intuitive kind. In the real numbers, all of the statements above should probably make sense due to past experience with the real number system.

In some of the arguments that follow we use the definition of function for the binary options described without explicitly saying so. For instance, it is understood that if $a+b = c$ and $a+b = d$ then we can conclude that $c = d$, meaning that c and d are the same element of S. This is because $+$ is a function, which means it must be true (by definition) that $+(a, b)$ is unique. This is important to understand even though it is convention not to mention it during arguments of this kind.

Definition 5

We write $a\frac{1}{b} = \frac{a}{b}$. We also write $a - b = a + -b$.

While our goal is to develop the field of real numbers, the axioms stated thus far do not do so. In fact, it is possible to have a field with only two elements.

Example 1.3. *Show the set consisting of* $\{0, 1\}$ *with operations* $0+0 = 0$, $0+1 = 1+0 = 1$, $1 + 1 = 0$, $(0)(0) = 0$, $(0)(1) = (1)(0) = 0$ *and* $(1)(1) = 1$ *is a field.*

Proof. We refer to this field as $\mathbb{Z}_2$. Identity, Commutativity and Associativity follow directly from the operation definitions. The multiplicative inverse of 1 is 1 and the additive inverse of 1 is 1. The additive inverse of 0 is 0. Since all axioms of a field are satisfied, $\mathbb{Z}_2$ is a field. □

As the example above demonstrates, not all fields are the real number system that we are trying to develop, and at this point we refrain from calling the elements of the field S "numbers."

We first make an observation that we use in these arguments without referencing the justification. That is, that it is true that whenever $a = b$ and a, b and c are elements of a field S, it follows that $a + c = b + c$ and that $ac = bc$. This is because addition and multiplication are binary operations, which means that they are functions. If a and b are the same element of S then since functions have a single output for each input (the image of a point is unique for a function), we know $+(a, c)$ is unique, meaning that $+(a, c) = +(b, c)$. Likewise $\cdot(a, c) = \cdot(b, c)$ since $\cdot$ is a function as well. It is common to not mention this when this property is used, so we just note it here, and after this point we will add or multiply both sides of an equation by an element of S and assume it is understood that doing so results in another true equation without further justification.

Likewise, in set theory all objects are sets (even if their elements are not specified) and two sets are equal if they have the same elements. Thus, if one set is equal to a second and a third it equal to the second then the first and third are equal. This justifies the (never mentioned) fact that if $a = b$ and $b = c$ then $a = c$.

Theorem 1.1. *Let S be a field. The identities $0, 1$ for S are unique. In other words, if $b \in S$ and $a + b = a$ for all $a \in S$ then $b = 0$, and if $b \in S$ and $ab = a$ for all $a \in S$ then $b = 1$.*

Proof. Suppose 0 and $0'$ are additive identities in S. Then $0 = 0 + 0' = 0'$ (by Identity). Hence, $0 = 0'$. Similarly, if $1, 1'$ are multiplicative identities then $1 = (1)(1') = 1'$. □

Theorem 1.2. *Let S be a field. For each $a \in S$ if $a + b = 0$ for some $b \in S$ then $b = -a$. If $a \neq 0$ and $ab = 1$ for some $b \in S$ then $b = \frac{1}{a}$.*

Proof. Let $a \in S$ and suppose that s, t are both additive inverses of a. Then $a+s = 0 = a+t$ (by Identity), so $-a+(a+t) = -a+(a+s)$, so $(-a+a)+s = (-a+a)+t$ (by Associativity) and $0 + s = 0 + t$ (by Inverses), so $s = t$ (by Identity).

Similarly, if s, t are multiplicative inverses of a non-zero point a then $\frac{1}{a}(as) = \frac{1}{a}(at)$, so $(\frac{1}{a}a)(s) = (\frac{1}{a}a)(t)$ and $1s = 1t$ so $s = t$. □

Note in the structure of the previous proof that we did not quote every axiom a second time in the latter part of the proof. It is fairly normal (and generally accepted) to not quote an axiom every time it is used if a similar application of the axiom has already been used within the argument. It is also common to not quote the axiom at all once it has been used enough times and the audience is understood to be one that is already aware of how the axiom is being used. We will continue to quote most uses of axioms for now.

Theorem 1.3. *Let S be a field. For any $a \in S$, $a(0) = 0$.*

Proof. We know $a(0) = a(0+0) = a(0) + a(0)$ (by Identity and Distributivity), so $-a(0) + a(0) = -a(0) + (a(0) + a(0)$, and $0 = (-a(0) + a(0)) + a(0)$ so $0 = 0 + a(0) = a(0)$.

□

Theorem 1.4. *Let S be a field. For any $a \in S$, $(-1)(a) = -a$*

Proof. Since $1+-1 = 0$, by Theorem 1.3 it follows that $0 = a(0) = a(1+-1) = a(1)+a(-1)$ (by Distributivity), so $a(-1)$ is the (unique) additive inverse of a and is therefore $-a$ by Theorem 1.2.

□

Theorem 1.5. *Let S be a field. For any $a, b \in S$, $(-b)(a) = -ba$*

Proof. By the Theorem 1.4, $(-b)(a) = ((-1)b)(a) = -1(ba) = -ba$ by Associativity.

□

Theorem 1.6. *Let a, b be non-zero elements of a field S. Then $\frac{1}{a}\frac{1}{b} = \frac{1}{ab}$.*

Proof. We know that $(\frac{1}{a}\frac{1}{b})(ab) = (a\frac{1}{a})(b\frac{1}{b})$ by Associativity and Commutativity, and this is equal to $(1)(1) = 1$, which means that $\frac{1}{a}\frac{1}{b}$ is the unique multiplicative inverse of ab which makes $\frac{1}{a}\frac{1}{b} = \frac{1}{ab}$. □

Theorem 1.7. *Let a, b be non-zero elements of a field S. Then $\frac{1}{b} + \frac{1}{a} = \frac{a+b}{ab}$.*

Proof. From the preceding theorem we know that $\frac{a+b}{ab} = (a+b)(\frac{1}{a}\frac{1}{b})$. By the Distributive, Associative and Commutative properties and Identity, this is $(a\frac{1}{a})\frac{1}{b} + (b\frac{1}{b})\frac{1}{a} = (1)(\frac{1}{a}) + (1)\frac{1}{b} = \frac{1}{b} + \frac{1}{a}$. □

Definition 6

A field S is *ordered* if there is a set $P \subseteq S$ of elements of S which are referred to as being *positive numbers*, satisfying the following conditions for all $a, b, c \in S$.

(a) Trichotomy: Exactly one of $a \in P$, $-a \in P$ or $a = 0$ is true for each $a \in S$.

(b) Closure: If $a, b \in P$ then $a + b \in P$ and $ab \in P$.

We use the notation $a < b$ or $b > a$ to mean $b - a \in P$ (note that $a > 0$ is thus the same as $a - 0 = a \in P$). The notation $a \leq b$ means that either $a = b$ or $a < b$. If $a < 0$ we refer to a as being *negative* or a *negative number*. We let 2 represent $1+1$, 3 denote $2+1$, and so on. Note that there are ordered fields which do not consist of numbers, but we are focused on the real line in this course so at this point we will use the term ”number” to refer to an element of an ordered field.

From this point forward we may omit referencing the standard field axioms (other than the order axioms) when we use them if their use seems fairly clear. Deciding which axioms or theorems should be explicitly stated in theorems depends on the audience and context and is difficult, particularly for readers beginning with proof writing. It is never wrong to include too much detail, specifying every axiom and theorem. So, when in doubt, it is sensible to write more than the reader needs to see.

For the theorems that follow in this chapter, it is understood that S is an ordered field and P is a subset of S consisting of the positive numbers satisfying the order axioms listed, and that all numbers stated are elements of S.

Theorem 1.8. *Let S be an ordered field in which $a < b$ and $c \in P$. Then $ca < cb$.*

Proof. Since $a < b$, $b - a \in P$ by definition, so $c(b - a) \in P$ by Closure, so $bc - ac \in P$, which means $bc > ac$.

□

Theorem 1.9. *Let S be an ordered field in which $a < b$ and $c \in S$. Then $a + c < b + c$.*

Proof. Since $a < b$, $b - a \in P$ so $(b + c) - (a + c) \in P$ by Closure.

□

Theorem 1.10. $1 > 0$.

Proof. By the Identity axiom, $1 \neq 0$. Suppose $1 < 0$. Then $0-1 = -1 \in P$, so $(-1)(-1) \in P$ by Closure, so $1 \in P$, which contradicts Trichotomy and is therefore impossible. Thus, by Trichotomy it must follow that $1 > 0$.

□

Theorem 1.11. *Let S be an ordered field in which $a < b$ and $-c \in P$. Then $ca > cb$.*

Proof. Since $-c \in P$, $-c(a) < -c(b)$ by Theorem 1.8. So, $-ca < -cb$ and $-ca + (ca + cb) < -cb + (ca + cb)$ by Theorem 1.9, so $cb < ca$.

□

Theorem 1.12. *Let S be an ordered field and let $a < b$ and $b < c$. Then $a < c$.*

Proof. Since $b - a \in P$ and $c - b \in P$ we know that $(b - a) + (c - b) = c - a \in P$ by Closure. □

Theorem 1.13. *Let S be an ordered field and let $a \neq 0$. Then a is positive if and only if $\frac{1}{a}$ is positive.*

Proof. Assume $a > 0$. Suppose $\frac{1}{a} < 0$. Then $a\frac{1}{a} < 0$ by Theorem 1.11 so $1 < 0$, a contradiction.

Likewise, if $\frac{1}{a}$ is positive then, since its multiplicative inverse is a, by the preceding argument a is positive. □

Definition 7

A subset A of an ordered field S is said to be *bounded above* if there is a point $u \in S$ so that $u \geq x$ for all $x \in A$. If this is the case then we call u an *upper bound* for A. Similarly, A is said to be *bounded below* if there is a point $l \in S$ so that $l \leq x$ for all $x \in A$. If this is the case then we call l a *lower bound* for A. A set is *bounded* if it has both an upper and a lower bound. If there is a point $t \in S$ which is an upper bound of A so that $t \leq u$ for every upper bound u of A then we call t the *least upper bound* or *supremum* of A, denoted $\sup(A)$ or $\sup A$. If there is a point $s \in S$ which is a lower bound of A so that $s \geq l$ for every lower bound l of A then we call s the *greatest lower bound* or *infimum* of A, denoted $\inf(A)$ or $\inf A$. An ordered field is said to be *complete* if every set which is bounded above has a least upper bound.

If there is a value $M \in A$ so that $x \leq M$ for every $x \in A$ then $M = \sup(A)$ and we say that $M = \max(A)$ (also written $\max A$) is the *maximum* or *largest* or *last* point of A.

If there is a value $m \in A$ so that $x \geq m$ for every $x \in A$ then $m = \inf(A)$ and we say that $m = \min(A)$ (also written $\min A$) is the *minimum* or *smallest* or *first* point of A.

We may remove parentheses or braces if the notation is less cumbersome. For instance, if A is a finite set $A = \{x_1, x_2, x_3, ..., x_n\}$ then we may write $\min(A) = \min\{x_1, x_2, ..., x_n\}$ or $\min(x_1, x_2, ..., x_3)$ instead of $\min(\{x_1, x_2, ..., x_n\})$.

We use the notation $\sup\limits_{x \in A} f(x)$ to mean the supremum of $f(A)$. If the variable is understood, we may instead write $\sup\limits_{A} f(x)$. Similarly, we use $\inf\limits_{x \in A} f(x)$ to mean the infimum of $f(A)$. If the variable is understood, we may instead write $\inf\limits_{A} f(x)$.

Definition 8

An *interval* is a set I such that if $a, b \in I$ and $a < x < b$ then $x \in I$. The *open interval* (a, b) will denote the set of points x satisfying $a < x < b$. For this text, we assume all open intervals listed are non-empty, and when the notation (a, b) is used it is implied that $a < b$. For $a \leq b$ we let the *closed interval* $[a, b] = \{x \in \mathbb{R} | a \leq x \leq b\}$ (and when we write $[a, b]$ it is implied that we are stating $a \leq b$). A half open interval is an open interval plus one of its two end points $[a, b) = \{x \in \mathbb{R} | a \leq x < b\}$ or $(a, b] = \{x \in \mathbb{R} | a < x \leq b\}$. An *extended interval* is one of: a *ray* (a set of the form $(a, \infty) = \{x \in \mathbb{R} | x > a\}, (-\infty, a), = \{x \in \mathbb{R} | x < a\}$, $[a, \infty) = \{x \in \mathbb{R} | x \geq a\}, (-\infty, a] = \{x \in \mathbb{R} | x \leq a\}$) or $\mathbb{R}$. In each of these expressions, a and b are referred to as *end points* of the interval. A ray containing its supremum or infimum is a *closed ray* and a ray which does not contain its supremum or infimum is an *open ray*.

We leave as an exercise to the reader (in the exercises for this chapter) to prove that a set I is an interval if and only if it is one of: a closed interval, an open interval, a half-open interval or an extended interval.

Example 1.4. *Let $A = [0, 1)$. Find:*
(a) A lower bound for A.
(b) The least upper bound of A
(c) The greatest lower bound of A.

Solution. (a) Any number less than or equal to zero would be a lower bound of A. For instance, -10 is a lower bound for A.
(b) The least upper bound of A is 1.
(c) The greatest lower bound of A is 0.

□

Definition 9

For any set $A \subseteq S$, the *coset* of A under multiplication by the number c is denoted by $cA = Ac = \{cx \in S | x \in A\}$. We also write $(-1)A$ as $-A$. The coset of A under addition by c is $A + c = c + A = \{c + x | x \in A\}$.

Notice that not all bounded sets in $\mathbb{R}$ have maxima or minima, but they all have suprema and infima.

Example 1.5. *Let $A = \{0, 1, 2\}$. Find:*
(a) $2A$
(b) $-A$
(c) $A + 5$

Solution. (a) $2A = \{0, 2, 4\}$.

(b) $-A = \{-2, -1, 0\}$.

(c) $A + 5 = \{5, 6, 7\}$.

□

The following axiom, referred to as the Completeness or Least Upper Bound axiom (or the Connectedness axiom) of the real numbers (depending on your audience) is that a set which has an upper bound always has a least upper bound. An ordered field which satisfies this axiom is called a *complete* ordered field (there is only one complete ordered field up to isomorphism or homeomorphism, but it is not necessary for us to prove that in this text). Assuming the completeness axiom is equivalent to assuming as an axiom that all decimals represent real numbers in the usual ordering and that the natural numbers are not bounded above.

When presenting calculus explanations to a calculus class, it may be more helpful to them to simply assume the aforementioned property that all decimals represent real numbers in the usual ordering and that the natural numbers are not bounded above and use this to describe why any set which is bounded above has a least upper bound. This is because in most calculus classes there isn't enough time to formally develop every theorem in calculus from fundamentals so you don't begin with teaching them about ordered fields and just pretend that the basics of the real number structure were handled in their algebra classes (which they were not, but it is convenient to proceed as if they were). If that is the assumption then one can essentially let $n_0.n_1n_2...n_k$ always be the largest decimal terminating at the kth place past the decimal that does not exceed all elements of a bounded set S for each natural number k, and then argue that $n_0.n_1n_2n_3...$ is the least upper bound for S.

For our development, however, we will assume the following as the axiom rather than the thing to be proven. Discussing why these two approaches are equivalent is addressed in the Supplementary Materials chapter (in the Decimals section) at the end of the single variable portion of the text for those who are interested.

Definition 10

We say an ordered field F is *complete* if every non-empty subset of F which is bounded above has a least upper bound.

We assume that there is a complete ordered field which we refer to as $\mathbb{R}$ as a final axiom.

Completeness Axiom: There is a complete ordered field $\mathbb{R}$.

All sets defined hereafter are assumed to be contained in $\mathbb{R}$ unless otherwise specified (when we reach the multivariable portion of the text we will assume sets to be contained in $\mathbb{R}^n$, but for now we assume they are sets of real numbers).

From this point on we may assume the standard properties that we have proven about an order field in the preceding theorems without referencing them, as long as their application seems clear.

Theorem 1.14. *If $A \subset \mathbb{R}$ and A has a lower bound, then A has a greatest lower bound.*

Proof. Let s be a lower bound for A. Then $-s \geq -x$ for every $x \in A$, which means that $-A$ is bounded above and has a least upper bound u. This means $u \geq -x$ and hence $-u \leq x$ for all $x \in A$, so $-u$ is a lower bound for A. If $-u < l$ then $u > -l$ which means that $-l$ is not an upper bound for $-A$, so there is some $-x \in -A$ so that $-l < -x$ and $l > x$, where $x \in A$. Thus, l is not a lower bound for A, which means that $-u$ is the greatest lower bound of A. □

Theorem 1.15. *Approximation Property. Let $A \subseteq \mathbb{R}$. If A is bounded above and $c < \sup(A)$ then there is some $a \in A$ so that $c < a \leq \sup(A)$. If A is bounded below and $c > \inf(A)$ then there is some $a \in A$ so that $\inf(A) \leq a < c$.*

Proof. First, assume A is bounded above. Since $\sup(A)$ is the least upper bound for A and $c < \sup(A)$ we know that c is not an upper bound for A, which means that there is some $a \in A$ so that $c < a$, and since $\sup(A)$ is an upper bound for A it follows that $c < a \leq \sup(A)$.

Next, assume A is bounded below. As before, since $\inf(A)$ is the greatest lower bound for A and $c > \inf(A)$ we know that c is not a lower bound for A, which means that there is some $a \in A$ so that $c > a$, and since $\inf(A)$ is a lower bound for A it follows that $\inf(A) \leq a < c$.

□

Definition 11

The *absolute value* of a is defined by setting $|a| = a$ if $a \geq 0$ and $|a| = -a$ if $a < 0$. We also define the *distance* from point p to point q to be $|p - q|$.

Theorem 1.16. *For any $a \in \mathbb{R}$, $-|a| \leq a \leq |a|$.*

The proof of this theorem is an exercise in the exercises below (and is proven in the solutions below). In various points in this text an exercise may be quoted as the exercise where it is proven or it may be listed as a theorem and then later proven in an exercised and we are not terribly consistent about which convention is used. Essentially, when seeing the statement of an exercise is important for understanding a theorem and we think the reader might have trouble filling in the missing details, or when a theorem is going to be used many times, we may state it as a theorem even if we defer the proof to an exercise. In cases where something has already been proven as an exercise before the first time it is used (or the result seems likely to be obvious), we are more likely to simply cite the exercise itself rather than restate it as a theorem.

Theorem 1.17. *Let $\epsilon > 0$ and $c \in \mathbb{R}$. Then, for every $x \in \mathbb{R}$,*
(a) $|x| < \epsilon$ if and only if $-\epsilon < x < \epsilon$
(b) $|x - c| < \epsilon$ if and only if $c - \epsilon < x < c + \epsilon$.
(c) $|x - c| \leq \epsilon$ if and only if $c - \epsilon \leq x \leq c + \epsilon$

Proof. (a) First, note that $-|x| \leq x \leq |x|$ by Theorem 1.16. If $|x| < \epsilon$ then $-\epsilon < -|x|$, so $-\epsilon < -|x| \leq x \leq |x| < \epsilon$. On the other hand, if $-\epsilon < x < \epsilon$ then if $x \geq 0$ then $-\epsilon < 0 \leq |x| = x < \epsilon$ and if $x < 0$ then $|x| = -x < \epsilon$, so $-\epsilon < x < 0 \leq |x| < \epsilon$ by Theorem 1.11.

(b) By part (a) we know that $|x - c| < \epsilon$ if and only if $-\epsilon < x - c < \epsilon$, which is true if and only if $c - \epsilon < x < c + \epsilon$.

(c) By part (b) we need only check the case where $|x - c| = \epsilon$, which happens if and only if $x - c = \epsilon$ or $x - c = -\epsilon$ by definition of absolute value. Thus, $|x - c| \leq \epsilon$ if and only if $|x - c| < \epsilon$ or $|x - c| = \epsilon$, which is true if and only if $c - \epsilon < x < c + \epsilon$ or $x = c - \epsilon$ or $x = c + \epsilon$, which is true if and only if $c - \epsilon \leq x \leq c + \epsilon$.

□

The preceding theorem tells us that we can think of the absolute value of a difference as being the same as the idea of distance, essentially. The statement "$|x - c| < \epsilon$" means the same thing as "the distance from x to c is less than ϵ."

Example 1.6. *Write $\{x \in \mathbb{R} | |x - 2| < 4\}$ as an interval.*

Solution: $(2 - 4, 2 + 4) = (-2, 6)$.

Theorem 1.18. *The Triangle Inequality. For any $a, b \in \mathbb{R}$:*
(i) $|a| + |b| \geq |a + b|$
(ii) $|a| - |b| \leq |a - b|$

Proof. (i) We see from Theorem 1.16 that $-|a| \leq a \leq |a|$ and $-|b| \leq b \leq |b|$, so $-(|a|+|b|) \leq a + b \leq (|a| + |b|)$ by Exercise 1.4. Thus, $|a + b| \leq |a| + |b|$ by Theorem 1.17.

(ii) By (i), $|b + (a - b)| \leq |b| + |a - b|$, so $|a| - |b| \leq |a - b|$.

□

The triangle inequality helps us to bound the distance between points if distances between intermediate points have known bounds. For instance, the following would be shown with the Triangle Inequality:

Example 1.7. *Let $a > 0$ and let $|a - b| < \dfrac{a}{2}$. Then prove $\dfrac{a}{2} < b < \dfrac{3a}{2}$ using the Triangle Inequality.*

Solution: By the Triangle Inequality $|b-0| \leq |a-0| + |a-b| < a + \frac{a}{2} = \frac{3a}{2}$. By the Triangle Inequality (second part), $|a| - |b| \leq |a-b|$, which means that $a - |b| < \frac{a}{2}$, so $\frac{a}{2} < |b| = b$.

Theorem 1.19. *A set A is bounded if and only if there is some $M > 0$ so that $-M \leq x \leq M$ for every $x \in A$, which is true if and only if $|x| < M$ for all $x \in A$.*

The proof is left as an exercise. We will use this theorem's result as an equivalent form of a set being "bounded" throughout the remainder of this text (without necessarily referencing this theorem).

Definition 12

We say a function $f : D \to \mathbb{R}$ is *bounded* if $ran(f)$ is bounded.

Example 1.8. *Let $f(x) = x^2$ and let $g(x) = 4 - x$ on $[0,3]$. Show that $f(x) + g(x) \leq 13$.*

Solution: First, if $x \in [0,3]$ then $x^2 \leq 3^2$ by Exercise 1.5, which means that $\sup_{[0,3]} f(x) = 9$. Also, if $0 \leq x < y$ then $-x > -y$ by Theorem 1.11 so $4 - x < 4 - y$ by Theorem 1.9. By Exercise 1.19, we know that $\sup_{[0,3]} f(x) + g(x) \leq 9 + 4 = 13$.

Exercises:

Exercise 1.1. *(DeMorgan's Laws) Let* $\{A_\alpha\}_{\alpha\in J}$ *be a collection of subsets of a set* X*, where* J *is an arbitrary indexing set. Then*

(a) $\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha) = X\setminus\bigcup_{\alpha\in J}A_\alpha$ *and*

(b) $\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha) = X\setminus\bigcap_{\alpha\in J}A_\alpha.$

In the proofs of the next two exercises, cite each axiom and theorem used in your proof.

Exercise 1.2. *Let* $a, b, c, d \in S$*, where* S *is a field, and let* $a, b \neq 0$*. Then* $\dfrac{c}{a}\dfrac{d}{b} = \dfrac{cd}{ab}$*.*

Exercise 1.3. *If* a, b *are non-zero elements of a field and* c, d *are elements of the field then* $\dfrac{c}{b} + \dfrac{d}{a} = \dfrac{ca + bd}{ab}$*.*

For the proofs of the remaining exercises in this section you are no longer required to cite the axioms of a field when they are used, as long as their application is clear in each step, but you should cite uses of the order and completeness axioms when these are used.

Exercise 1.4. *Let* $a < b$ *and let* $c < d$*. Then* $a + c < b + d$*.*

Exercise 1.5. *Let* $0 \le a < b$ *and let* $0 \le c < d$*. Then* $ac < bd$*.*

Exercise 1.6. *An ordered field has no smallest positive element.*

Exercise 1.7. *Let* F *be an ordered field. If* $a \in F$ *then* $a^2 \ge 0$*.*

Exercise 1.8. *Let* F *be an ordered field and let* $0 < a < b$*, for some* $a, b \in F$*. Then* $0 < \dfrac{1}{b} < \dfrac{1}{a}$*.*

Exercise 1.9. *Prove that, for any* $a \in \mathbb{R}$*,* $-|a| \le a \le |a|$*.*

Exercise 1.10. *Let* F *be an ordered field and let* $a, b \in F$*. Then* $|ab| = |a||b|$*.*

Exercise 1.11. *Let F be an ordered field and let $a \in F$. If $|a| < \epsilon$ for every $\epsilon > 0$ in F, then $a = 0$.*

Exercise 1.12. *If $A \subset \mathbb{R}$ which is bounded above then for any $c > 0$, the set cA has a least upper bound $c\sup(A)$, and the set $-cA$ has a greatest lower bound $-c\sup(A)$.*

Exercise 1.13. *Prove that if $A \subset \mathbb{R}$ which is bounded below then for any $c > 0$, the set cA has a greatest lower bound $c\inf(A)$, and the set $-cA$ has a least upper bound $-c\inf(A)$*

Exercise 1.14. *Prove that a set A is bounded if and only if there is some $M > 0$ so that $-M \leq x \leq M$ for every $x \in A$, which is true if and only if $|x| \leq M$ for all $x \in A$.*

Exercise 1.15. *If x, y are real numbers so that for every $\epsilon > 0$ it is true that $|x - y| < \epsilon$ then $x = y$.*

Exercise 1.16. *Let $A, B \subseteq \mathbb{R}$ be non-empty sets.*

(a) If B is bounded above and, for each $x \in A$, there is a point $y \in B$ so that $y \geq x$. Then $\sup(A) \leq \sup(B)$.

(b) If B is bounded below and, for each $x \in A$, there is a $y \in B$ so that $y \leq x$ then $\inf(A) \geq \inf(B)$.

Exercise 1.17. *Let A and B be non-empty sets with $A \subseteq B$.*

(a) If B is bounded above then $\sup(A) \leq \sup(B)$.

(b) If B is bounded below then $\inf(B) \leq \inf(A)$.

Exercise 1.18. *Let $A \subset \mathbb{R}$. If A is bounded above then $\sup(A + k) = \sup(A) + k$. If A is bounded below then $\inf(A + k) = \inf(A) + k$.*

Exercise 1.19. *Let $f, g : [c, d] \to \mathbb{R}$ be bounded functions. Then $\sup\limits_{x\in[c,d]} f(x) + \sup\limits_{x\in[c,d]} g(x) \geq \sup\limits_{x\in[c,d]} f(x) + g(x)$ and $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} f(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$. Likewise, $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} g(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$.*

Exercise 1.20. *Let $f, g : E \to \mathbb{R}$ be bounded. Then $f(x)g(x)$ is bounded.*

Exercise 1.21. *Let $I \subseteq \mathbb{R}$. Then I is an interval if and only if I is one of the following: an open interval, a closed interval, a half-open interval or an extended interval.*

Hints:

Hint to Exercise 1.1. *(DeMorgan's Laws) Let $\{A_\alpha\}_{\alpha\in J}$ be a collection of subsets of a set X, where J is an arbitrary indexing set. Then*

(a) $\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcup_{\alpha\in J}A_\alpha$ *and*

(b) $\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcap_{\alpha\in J}A_\alpha.$

Write down the definition of what it means for x to be in the left and right sides of the equations, respectively.

In each case, let x be an element of the set in the left side of the equation. Explain why that means x is an element of the set on the right side of the equation. Then let x be an arbitrary point in the set on the right side of the equation and explain why that means x is in the set on the left side of the equation.

Hint to Exercise 1.2. *Let $a,b,c,d\in S$, where S is a field, and let $a,b\neq 0$. Then $\dfrac{c}{a}\dfrac{d}{b}=\dfrac{cd}{ab}$.*

Remember that the definition of $\dfrac{c}{b}$ is c times the multiplicative inverse of b. Also recall that it has already been shown that $\dfrac{1}{a}\dfrac{1}{b}=\dfrac{1}{ab}$. Write down what $\dfrac{cd}{ab}$ means and look through the field axioms.

Hint to Exercise 1.3. *If a,b are non-zero elements of a field and c,d are elements of the field then $\dfrac{c}{b}+\dfrac{d}{a}=\dfrac{ca+bd}{ab}$.*

Remember that we have already shown that $\dfrac{1}{b}+\dfrac{1}{a}=\dfrac{a+b}{ab}$. Try to use field axioms and the definition of $\dfrac{ca+bd}{ab}$.

Hint to Exercise 1.4. *Let $a<b$ and let $c<d$. Then $a+c<b+d$.*

First, show that $a+c<b+c$ and then show that $b+c<b+d$.

Hint to Exercise 1.5. *Let $0\leq a<b$ and let $0\leq c<d$. Then $ac<bd$.*

Use Theorem 1.8.

Hint to Exercise 1.6. *An ordered field has no smallest positive element.*

You might want to start by proving that $0<\dfrac{1}{2}<1$.

Hint to Exercise 1.7. *Let F be an ordered field. If $a \in F$ then $a^2 \geq 0$.*

Try breaking the problem down into cases. Trichotomy says $a > 0$ or $a = 0$ or $a < 0$. Prove the result in each case separately.

Hint to Exercise 1.8. *Let F be an ordered field and let $0 < a < b$, for some $a, b \in F$. Then $0 < \frac{1}{b} < \frac{1}{a}$.*

See if you can find a positive number to multiply the terms in the inequality $0 < a < b$ by to arrive at $0 < \frac{1}{b} < \frac{1}{a}$.

Hint to Exercise 1.9. *Prove that, for any $a \in \mathbb{R}$, $-|a| \leq a \leq |a|$.*

Try looking at the problem in cases. By Trichotomy, $a > 0$ or $a < 0$ or $a = 0$.

Hint to Exercise 1.10. *Let F be an ordered field and let $a, b \in F$. Then $|ab| = |a||b|$.*

Try looking at the problem in cases. By Trichotomy, $a > 0$ or $a < 0$ or $a = 0$, and $b > 0$ or $b < 0$ or $b = 0$.

Hint to Exercise 1.11. *Let F be an ordered field and let $a \in F$. If $|a| < \epsilon$ for every $\epsilon > 0$ in F, then $a = 0$.*

Suppose $a \neq 0$. What can be said about $|a|$? Does this contradict the hypothesis?

Hint to Exercise 1.12. *If $A \subset \mathbb{R}$ which is bounded above then for any $c > 0$, the set cA has a least upper bound $c\sup(A)$, and the set $-cA$ has a greatest lower bound $-c\sup(A)$.*

First, take an element $x \in A$ and explain why $cx \leq c\sup(A)$. Then, suppose u is an upper bound for cA and explain why $\frac{u}{c}$ is an upper bound for A. Use this to explain why $c\sup(A)$ is the least upper bound of cA. Then do something similar for the other half of the theorem.

Hint to Exercise 1.13. *Prove that if $A \subset \mathbb{R}$ which is bounded below then for any $c > 0$, the set $cA = \{cx \in \mathbb{R} | a \in A\}$ has a greatest lower bound $c\inf(A)$, and the set $-cA$ has a least upper bound $-c\inf(A)$*

The argument is very similar to that described for the proof of the preceding exercise.

Hint to Exercise 1.14. *Prove that a set A is bounded if and only if there is some $M > 0$ so that $-M \leq x \leq M$ for every $x \in A$, which is true if and only if $|x| \leq M$ for all $x \in A$.*

Write down the definition of what it means for a set to be bounded. Explain why this means the absolute values of the elements of the set are bounded. Then explain why the absolute values of elements of a set being bounded would imply that the set is bounded. You might also consider using Theorem 1.16.

Hint to Exercise 1.15. *If x, y are real numbers so that for every $\epsilon > 0$ it is true that $|x-y| < \epsilon$ then $x = y$.*

You could use Theorem 1.17, or methods similar to the proof of that theorem.

Hint to Exercise 1.16. *Let $A, B \subseteq \mathbb{R}$ be non-empty sets.*

(a) If B is bounded above and, for each $x \in A$, there is a point $y \in B$ so that $y \geq x$. Then $\sup(A) \leq \sup(B)$.

(b) If B is bounded below and, for each $x \in A$, there is a $y \in B$ so that $y \leq x$ then $\inf(A) \geq \inf(B)$.

Start with an upper bound for B and explain why that is also an upper bound for A. For the second part, start with a lower bound for B and explain why that is also a lower bound for A.

Hint to Exercise 1.17. *Let A and B be non-empty sets with $A \subseteq B$.*

(a) If B is bounded above then $\sup(A) \leq \sup(B)$.

(b) If B is bounded below then $\inf(B) \leq \inf(A)$.

Show that an upper bound for B is an upper bound for A and that a lower bound for B is a lower bound for A, and then explain why this implies the conclusion (consider the definitions of least upper bound and greatest lower bound).

Hint to Exercise 1.18. *Let $A \subset \mathbb{R}$ and $k \in \mathbb{R}$ and define $A + k = \{x + k \in \mathbb{R} | a \in A\}$. If A is bounded above then $\sup(A+k) = \sup(A) + k$. If A is bounded below then $\inf(A+k) = \inf(A) - k$.*

Try to parallel the strategy of Exercise 1.12 somewhat, using addition and subtraction instead of multiplication and division.

Hint to Exercise 1.19. *Let $f, g : [c,d] \to \mathbb{R}$ be bounded functions. Then $\sup\limits_{x\in[c,d]} f(x) + \sup\limits_{x\in[c,d]} g(x) \geq \sup\limits_{x\in[c,d]} f(x) + g(x)$ and $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} f(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$. Likewise, $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} g(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$.*

Try to explain why $f(z) + g(z) \geq \inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} g(x)$ for each $z \in [c,d]$.

Hint to Exercise 1.20. *Let $f, g : E \to \mathbb{R}$ be bounded. Then $f(x)g(x)$ is bounded.*

You could use Theorem 1.19.

Hint to Exercise 1.21. *Let $I \subseteq \mathbb{R}$. Then I is an interval if and only if I is one of the following: an open interval, a closed interval, a half-open interval or an extended interval.*

Start out with the definitions of each type of interval listed and explain why they satisfy the definition of being an interval, which is that an interval is a set containing every point which is between any two elements of that set. Then, use the definition of an interval and break down the possible cases of the interval being bounded above, bounded below, both above and below and neither bounded above nor below, and explain why these cases each imply that an interval is one of the listed types of interval.

Solutions:

Solution to Exercise 1.1. *(DeMorgan's Laws) Let $\{A_\alpha\}_{\alpha\in J}$ be a collection of subsets of a set X, where J is an arbitrary indexing set. Then*

(a) $\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcup_{\alpha\in J}A_\alpha$ *and*

(b) $\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcap_{\alpha\in J}A_\alpha$.

Proof. (a) Let $x\in\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha)$. Then for every $\alpha\in J$ it follows that $x\in X$ and $x\notin A_\alpha$ which means that $x\in X\setminus\displaystyle\bigcup_{\alpha\in J}A_\alpha$.

Let $x\in X\setminus\displaystyle\bigcup_{\alpha\in J}A_\alpha$. This means that $x\in X$ and x is not in any A_α which means that $x\in X\setminus A_\alpha$ for every $\alpha\in J$ and thus $x\in\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha)$.

Hence, $\displaystyle\bigcap_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcup_{\alpha\in J}A_\alpha$.

(b) Let $x\in\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha)$. Then for some $\beta\in J$ it follows that $x\in X$ and $x\notin A_\beta$. This means that $x\notin\displaystyle\bigcap_{\alpha\in J}A_\alpha$ and thus $x\in X\setminus\displaystyle\bigcap_{\alpha\in J}A_\alpha$.

Let $x\in X\setminus\displaystyle\bigcap_{\alpha\in J}A_\alpha$. Then $x\in X$ and since $x\notin\displaystyle\bigcap_{\alpha\in J}A_\alpha$, there is some $\beta\in J$ so that $x\notin A_\beta$. Thus, $x\in X\setminus A_\beta$ which means that $x\in\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha)$.

Thus, $\displaystyle\bigcup_{\alpha\in J}(X\setminus A_\alpha)=X\setminus\bigcap_{\alpha\in J}A_\alpha$. □

Solution to Exercise 1.2. *Let $a,b,c,d\in S$, where S is a field, and let $a,b\neq 0$. Then* $\dfrac{c}{a}\dfrac{d}{b}=\dfrac{cd}{ab}$.

Proof. Since we have already shown that $\dfrac{1}{a}\dfrac{1}{b}=\dfrac{1}{ab}$, we simply observe that $\dfrac{c}{a}\dfrac{d}{b}=(c\dfrac{1}{a})(d\dfrac{1}{b})$ by definition, which is $(cd)(\dfrac{1}{a}\dfrac{1}{b})=(cd)\dfrac{1}{ab}$ by the Associative and Commutative properties, which is the same as $\dfrac{cd}{ab}$ by definition. □

Solution to Exercise 1.3. *If a,b are non-zero elements of a field and c,d are elements of the field then* $\dfrac{c}{b}+\dfrac{d}{a}=\dfrac{ca+bd}{ab}$.

Proof. We know that $ca+bd=ca+bd$, so multiplying both sides by $\dfrac{1}{a}\dfrac{1}{b}$ gives $\dfrac{1}{a}\dfrac{1}{b}(ca+bd)=\dfrac{1}{a}\dfrac{1}{b}(ca+bd)$, and since we have shown that $\dfrac{1}{a}\dfrac{1}{b}=\dfrac{1}{ab}$ we can rewrite this as $\dfrac{1}{a}\dfrac{1}{b}(ca)+$

$\frac{1}{a}\frac{1}{b}(db) = \frac{ca+bd}{ab}$. By Commutativity and Associativity we can simplify the left side of this equation fo $(c\frac{1}{b})(a\frac{1}{a}) + (d\frac{1}{a})(b\frac{1}{b})$. By Inverses and Identity this further simplifies to $\frac{c}{b} + \frac{d}{a}$. Thus, $\frac{c}{b} + \frac{d}{a} = \frac{ca+bd}{ab}$. □

Solution to Exercise 1.4. *Let $a < b$ and let $c < d$. Then $a + c < b + d$.*

Proof. By Theorem 1.9 we know $a + c < b + c$. By the same theorem, $b + c < b + d$. Thus, by Theorem 1.12, we know that $a + c < b + d$. □

Solution to Exercise 1.5. *Let $0 \leq a < b$ and let $0 \leq c < d$. Then $ac < bd$.*

Proof. If $a = 0$ or $c = 0$ then $ac = 0$ and since $c, d > 0$ we know that $cd > 0$ by Closure, so $ac < bd$. Otherwise $a, c > 0$, in which case we know $ca < cb$ and $bc < bd$ by Theorem 1.8, which means that $ac < bd$ by Theorem 1.12. □

Solution to Exercise 1.6. *An ordered field has no smallest positive element.*

Proof. First, note that since we have shown $1 > 0$, it follows that $1+1 > 1$ so $2 > 1$ and 2 is positive. Hence, it follows from Theorem 1.13 that $\frac{1}{2}$ is positive, which means that $2(\frac{1}{2}) > 1(\frac{1}{2})$ so $\frac{1}{2} < 1$. Let $a > 0$. Then since $0 < \frac{1}{2} < 1$ we know that $0(a) < a(\frac{1}{2}) < a(1)$, so $\frac{a}{2}$ is a positive number which is less than a. Thus, there is no smallest positive number. □

Solution to Exercise 1.7. *Let F be an ordered field. If $a \in F$ then $a^2 \geq 0$.*

Proof. If $a > 0$ then $(a)(a) > 0$ by the Closure axiom for the positive numbers. If $a = 0$ then $(a)(a) = 0$. If $a < 0$ then $-a > 0$ so $(-a)(-a) > 0$ by Closure again, so $(-a)(-a) = (-1)(-1)(a)(a) = (1)(a^2) = a^2 > 0$. By Trichotomy, these are the only possible cases, so $a^2 \geq 0$ for every $a \in F$. □

Solution to Exercise 1.8. *Let F be an ordered field and let $0 < a < b$, for some $a, b \in F$. Then $0 < \frac{1}{b} < \frac{1}{a}$.*

Proof. Since a, b are positive, $ab > 0$ by the Closure axiom for the positive numbers. Thus, by Theorem 1.13, we know that $\frac{1}{ab} > 0$. Also, we have shown that $\frac{1}{ab} = \frac{1}{a}\frac{1}{b}$. Thus, since $a < b$ it follows that $\frac{1}{a}\frac{1}{b}a < \frac{1}{a}\frac{1}{b}b$ which means $\frac{1}{b} < \frac{1}{a}$. Finally, since multiplying two positive numbers always results in a positive number by Closure, we know that $0 < \frac{1}{b}$ so $0 < \frac{1}{b} < \frac{1}{a}$. □

Solution to Exercise 1.9. *Prove that for any $a \in \mathbb{R}$. $-|a| \leq a \leq |a|$.*

Proof. If $a \geq 0$ then $|a| = a \geq 0 \geq -a = -|a|$. If $a < 0$ then $|a| = -a > 0 > a = -|a|$. Thus, in each possible case the theorem is true. □

Solution to Exercise 1.10. *Let F be an ordered field and let $a, b \in F$. Then $|ab| = |a||b|$.*

Proof. If $a \geq 0$ and $b \geq 0$ then $|a||b| = ab = |ab|$. If $a \geq 0$ and $b < 0$ then $|a||b| = a(-b) = -ab = |ab|$. If $a < 0$ and $b < 0$ then $|ab| = ab = (-a)(-b) = |a||b|$. All possible cases are addressed in these three since the case where $a < 0$ and $b \geq 0$ is handled in the second case by renaming a and b, so this completes the proof. □

Solution to Exercise 1.11. *Let F be an ordered field and let $a \in F$. If $|a| < \epsilon$ for every $\epsilon > 0$ in F, then $a = 0$.*

Proof. If $a > 0$ then $|a| = a > 0$ which is impossible since $|a| < a$ by assumption. If $a < 0$ then $|a| = -a > 0$ which is, again, impossible, since $|a| < -a$ by assumption. Hence, $a = 0$. □

Solution to Exercise 1.12. *If $A \subset \mathbb{R}$ which is bounded above then for any $c > 0$, the set cA has a least upper bound $c\sup(A)$, and the set $-cA$ has a greatest lower bound $-c\sup(A)$.*

Proof. For every $x \in A$ we know that $x \leq \sup(A)$, so $cx \leq c\sup(A)$, which is an upper bound for cA. If u is an upper bound for cA then for every $x \in cA$ it must follow that $\frac{x}{c} \leq \frac{u}{c}$ so $\frac{u}{c}$ is an upper bound for A. This means that $\frac{u}{c} \geq \sup(A)$ and therefore $u \geq c\sup(A)$, so $c\sup(A)$ is the least upper bound of cA.

For every $x \in A$ we know that $x \leq \sup(A)$, so $-cx \geq -c\sup(A)$, which is a lower bound for $-cA$. If l is a lower bound for $-cA$ then for every $x \in A$ we know that $-cx \in -cA$ so $-cx \geq l$, which means that $x \leq \frac{l}{-c}$ so $\frac{l}{-c}$ is an upper bound for A. This means that $\frac{l}{-c} \geq \sup(A)$ and therefore $l \leq -c\sup(A)$, so $-c\sup(A)$ is the greatest lower bound of $-cA$.

□

Solution to Exercise 1.13. *Prove that if $A \subset \mathbb{R}$ which is bounded below then for any $c > 0$, the set cA has a greatest lower bound $c\inf(A)$, and the set $-cA$ has a least upper bound $-c\inf(A)$.*

Proof. For every $x \in A$ we know that $x \geq \inf(A)$, so $cx \geq c\inf(A)$, which is a lower bound for cA. If l is a lower bound for cA then for every $x \in cA$ it must follow that $\frac{x}{c} \geq \frac{l}{c}$ so $\frac{l}{c}$ is a lower bound for A. This means that $\frac{l}{c} \leq \inf(A)$ and therefore $l \leq c\inf(A)$, so $c\inf(A)$ is the greatest lower bound of cA.

For every $x \in A$ we know that $x \geq \inf(A)$, so $-cx \leq -c\inf(A)$, which is an upper bound for $-cA$. If u is an upper bound for $-cA$ then for every $x \in A$ we know that $-cx \in -cA$ so $-cx \leq u$, so $x \geq \frac{u}{-c}$ so $\frac{u}{-c}$ is a lower bound for A. This means that $\frac{u}{-c} \leq \inf(A)$ and therefore $u \geq -c\inf(A)$, so $-c\inf(A)$ is the least upper bound of $-cA$.

□

Solution to Exercise 1.14. *Prove that a set A is bounded if and only if there is some $M > 0$ so that $-M \leq x \leq M$ for every $x \in A$, which is true if and only if $|x| \leq M$ for all $x \in A$.*

Proof. First, let $M > 0$. Then $|x| \leq M$ for each $x \in A$ if and only if $-M \leq x \leq M$ for each $x \in A$ by Theorem 1.17.

Assume that A is bounded. Then there are bounds a, b for A so that $a \leq x \leq b$ for each $x \in A$. Let $M = \max(|a|, |b|)$. By Theorem 1.16, for each $x \in A$ it is true that $-M \leq |a| \leq a \leq x \leq b \leq |b| \leq M$, which means that $|x| \leq M$.

If M is a positive number so that $-M \leq x \leq M$ for all $x \in A$ then $-M$ is a lower bound for A and M is an upper bound for A, so A is bounded.

□

Solution to Exercise 1.15. *If x, y are real numbers so that for every $\epsilon > 0$ it is true that $|x - y| < \epsilon$ then $x = y$.*

Proof. By an earlier exercise, $|x - y| = 0$. If $x - y > 0$ then $|x - y| = x - y > 0$ and if $y - x > 0$ then $|x - y| = y - x > 0$. Thus, it is false that $x - y$ is positive or negative, so by Trichotomy we conclude that $x - y = 0$, so $x = y$. □

Solution to Exercise 1.16. *Let $A, B \subseteq \mathbb{R}$ be non-empty sets.*

(a) If B is bounded above and, for each $x \in A$, there is a point $y \in B$ so that $y \geq x$. Then $\sup(A) \leq \sup(B)$.

(b) If B is bounded below and, for each $x \in A$, there is a $y \in B$ so that $y \leq x$ then $\inf(A) \geq \inf(B)$.

Proof. (a) Let $u = \sup(B)$. Then for each $a \in A$ we know that there there is some $b \in B$ so that $a \leq b \leq u$ which means that $a \leq u$, so u is an upper bound for A. Hence, $\sup(A) \leq u$.

(b) Let $l = \inf(B)$. Then for each $a \in A$ there is some $b \in B$ so that $l \leq b \leq a$, which means that l is a lower bound for A. Hence, $l \leq \inf(A)$. □

Solution to Exercise 1.17. *Let A and B be non-empty sets with $A \subseteq B$.*

(a) If B is bounded above then $\sup(A) \leq \sup(B)$*.*

(b) If B is bounded below then $\inf(B) \leq \inf(A)$*.*

Proof. (a) For each $x \in A$ we know that $x \in B$, which means that $x \leq \sup(B)$. Thus, $\sup(B)$ is an upper bound for A, so $\sup(A) \leq \sup(B)$.

(b) For each $x \in A$ we know that $x \in B$, which means that $x \geq \inf(B)$. Thus, $\inf(B)$ is a lower bound for A, so $\inf(A) \geq \inf(B)$.

□

Solution to Exercise 1.18. *Let $A \subset \mathbb{R}$ and let $A+k = \{x+k \in \mathbb{R} | a \in A\}$. If A is bounded above then* $\sup(A+k) = \sup(A) + k$*. If A is bounded below then* $\inf(A+k) = \inf(A) + k$*.*

Proof. Let $x \in A$. Then $x \leq \sup(A)$ so $x + k \leq \sup(A) + k$, which is an upper bound of $A + k$. Let u be an upper bound of $A + k$. Then for any $x \in A$ we know that $x + k \leq u$ so $x \leq u - k$, which is an upper bound for A. Thus, $\sup(A) \leq u - k$ so $\sup(A) + k \leq u$, making $\sup(A) + k$ the least upper bound of $A + k$.

Next, if $x \in A$ then $x \geq \inf(A)$ so $x + k \geq \inf(A) + k$, which is a lower bound of $A + k$. Let l be an upper bound of $A + k$. Then for any $x \in A$ we know that $x + k \geq l$ so $x \geq l - k$, which is a lower bound for A. Thus, $\inf(A) \geq l - k$ so $\inf(A) + k \geq l$, making $\inf(A) + k$ the greatest lower bound of $A + k$. □

Solution to Exercise 1.19. *Let $f, g : [c, d] \to \mathbb{R}$ be bounded functions. Then* $\sup\limits_{x\in[c,d]} f(x) + \sup\limits_{x\in[c,d]} g(x) \geq \sup\limits_{x\in[c,d]} f(x) + g(x)$ *and* $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} f(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$*. Likewise,* $\inf\limits_{x\in[c,d]} f(x) + \inf\limits_{x\in[c,d]} g(x) \leq \inf\limits_{x\in[c,d]} f(x) + g(x)$*.*

Proof. Let $S_f = \inf\limits_{x\in[c,d]} f(x)$ and $S_g = \inf\limits_{x\in[c,d]} g(x)$. For every $x \in [c, d]$ we know that $f(x) \leq S_f$ and $g(x) \leq S_g$, which means that $f(x) + g(x) \leq S_f + S_g$, which is an upper bound for $\{f(x) + g(x) | x \in [c, d]\}$. This means that $S_f + S_g$ is greater than or equal to the least lower bound of $\{f(x) + g(x) | x \in [c, d]\}$, which is $\sup\limits_{x\in[c,d]} f(x) + g(x)$.

Let $I_f = \inf\limits_{x\in[c,d]} f(x)$ and $I_g = \inf\limits_{x\in[c,d]} g(x)$. For every $x \in [c, d]$ we know that $f(x) \geq I_f$ and $g(x) \geq I_g$, which means that $f(x) + g(x) \geq I_f + I_g$, which is a lower bound for $\{f(x) + g(x) | x \in [c, d]\}$. This means that $I_f + I_g$ is less than or equal to the greatest lower bound of $\{f(x) + g(x) | x \in [c, d]\}$, which is $\inf\limits_{x\in[c,d]} f(x) + g(x)$. □

Solution to Exercise 1.20. *Let $f, g : E \to \mathbb{R}$ be bounded. Then $f(x)g(x)$ is bounded.*

Proof. Since f and g are bounded, there are numbers M_f, M_g so that $|f(x)| \leq M_f$ and $|g(x)| \leq M_g$ for all $x \in E$. Hence, $|f(x)g(x)| = |f(x)||g(x)| \leq M_f M_g$ for all $x \in E$, which means that fg is bounded on E. □

Solution to Exercise 1.21. *Let $I \subseteq \mathbb{R}$. Then I is an interval if and only if I is one of the following: an open interval, a closed interval, a half-open interval or an extended interval.*

Proof. First, assume that I is an open or closed interval with end points $a \leq b$. If $a = b$ then I consists of at most one point and is an interval vacuously. Otherwise, if $c, d \in I$ with $c < d$ then $a \leq c < d \leq b$, so if $c < x < d$ then $a < x < b$ which means $x \in I$ by definition of open and closed interval. Next, assume that I is a ray with end point a. If $I = (a, \infty)$ or $[a, \infty)$ and $a \leq c < d$ and $c < x < d$ then again, $x \in I$ by definition since $a < x$. Similarly, if $I = (-\infty, a)$ or $I = (-\infty, a]$ then I is an interval since if $c < x < d \leq a$ then $x < a$ so $x \in I$. Finally, if $I = \mathbb{R}$ then all real numbers are in I so I is an interval.

Next, assume that I is an interval. Then if I is neither bounded above nor below then for any real x there are points $c, d \in I$ so that $c < x < d$ so $x \in I$, which means that $I = \mathbb{R}$. If I is bounded below but not above then let $a = \inf(I)$. If $x > a$ then x is not a lower bound for I, which means that for some point $c < x$ it is true that $c \in I$ by the Approximation Property. Since I is not bounded above there is some $d > x$ so that $d \in I$, and hence $c < x < d$, so $x \in I$. Thus, I is either $[a, \infty)$ or (a, ∞) depending on whether or not $a \in I$.

By similar reasoning, if I is bounded above but not below then let $a = \sup(I)$. If $x < a$ then x is not an upper bound for I, which means that for some point $d > x$ it is true that $d \in I$. Since I is not bounded below there is some $c < x$ with $c \in I$, and hence $c < x < d$, so $x \in I$. Thus, I is either $(-\infty, a)$ or $(-\infty, a]$ depending on whether or not $a \in I$.

If I is bounded above and below, let $a = \inf(I)$ and let $b = \sup(I)$. For any point $x \in (a, b)$ we can find $c \in (a, x)$ and $d \in (x, b)$ so $x \in I$ which means that $(a, b) \subseteq I$. Thus, $I = [a, b]$ or $[a, b)$ or $(a, b]$ or (a, b).

□

Chapter 2

Induction

The *natural numbers* are denoted by $\mathbb{N}$ and are the counting numbers $\{1, 2, 3, ...\}$ but using that as a definition before establishing induction is potentially problematic because it is not clear that such listings with three dots at the end are a valid way to create a set without induction. Because we are focused on the main results of advanced calculus in this text we will assume results about the natural numbers in this section without their proofs.

The Supplementary Materials chapter contains a development of the natural numbers without assuming any additional axioms. If the reader has time, it is recommended that the proofs of these properties be studied from the Supplementary Materials chapter.

Essentially, if this text is used for a math education major course and the goal is to finish a fair bit of the integration chapter in one semester then it is probably better to proceed as outlined below. If this text is used as a math major text and is the first proof course a student encounters with the goal of finishing Chapter 5 with a brief introduction to integration then foundations are probably more important and integration will be covered in a later course, so it might be wise to prove the theorems in the development in the supplementary materials.

Properties of $\mathbb{N}$ (and the theorem in the Supplementary Materials chapter where a justification can be found for each):

1 is the least element of $\mathbb{N}$ (by Theorem 7.1)

$\mathbb{N}$ is well-ordered, meaning that every non-empty subset of $\mathbb{N}$ has a least element. (by Theorem 7.3)

For every $n \in \mathbb{N}$, $n + 1 \in \mathbb{N}$. (by Theorem 7.1)

If $n \in \mathbb{N}$ there are no natural numbers between n and $n+1$ or between n and $n-1$. (by Theorem 7.3)

If $n > 1$ and $n \in \mathbb{N}$ then $n - 1 \in \mathbb{N}$ (by Theorem 7.1)

Theorem 2.1. *Principle of Mathematical Induction. Let $P(n)$ be a statement so that for each $n \in \mathbb{N}$ the following two statements are true:*

(a) $P(1)$ is true

and

(b) If $P(k)$ is true for some $k \in \mathbb{N}$ then $P(k + 1)$ is true

Then it follows that $P(n)$ is true for all $n \in \mathbb{N}$.

Proof. Let $S = \{n \in \mathbb{N} | P(n) \text{ is false }\}$. Suppose $S \neq \emptyset$. Then since $\mathbb{N}$ is well-ordered, there is a first element $m \in S$. We know $P(1)$ is true and since $1 \neq m$ and 1 is the least natural number, it must follow that $m > 1$. Thus, we know that $m - 1 \in \mathbb{N}$ and since m is the least element of S it follow that $m - 1 \notin S$, so $P(m - 1)$ is true. But then, by (b) we know that $P(m - 1 + 1)$ is true, so $P(m)$ is a true statement, contradicting $m \in S$. Hence, it follows that S is empty, so $P(n)$ is true for all natural numbers n. □

Note that the assumption that $P(k)$ is true in induction arguments is often referred to as the "induction hypothesis." The principle of mathematical induction feels like it should be true because if a first statement is true and it is true that whenever a given statement is true then the next statement it true then the second statement is true because the first is true, and the third is true because the second is true, and the fourth is true because the third is true and so on, so the statement is true for all natural numbers. Unfortunately, the "and so on" part of that discussion is the assumption of mathematical induction. Thus, while that description is not mathematically rigorous, it is helpful to assist some people with their intuition. Some people studying induction for the first time think "doesn't the assumption that the statement is true for a given k which is made in most induction proofs assume what we are trying to prove?" It does not, because we are only making such an assumption for an arbitrary k and using that to prove the statement would then be true for the $k + 1$st statement. This then shows that if the statement is true for a given k then it is true for $k + 1$. Most mathematical theorems are phrased like this. You prove something is true assuming certain hypotheses. No one is asserting the hypotheses are true, only that if they are true then a conclusion follows. When we make the induction hypothesis assumption we are not stating that we think $P(k)$ is true. We are demonstrating that if it is given that $P(k)$ is true then that would imply that $P(k + 1)$ is true as well.

Here is an example of a proof using induction. Since we may use the result later, we will label it as a theorem rather than an example.

Theorem 2.2. *For all natural numbers n, the sum $1 + 2 + ... + n = \frac{n(n+1)}{2}$.*

Proof. Proceeding by induction, when $n = 1$ we know that $1 = \frac{1(1+1)}{2}$ is true. Assuming the statement is true when $n = k \in \mathbb{N}$ we have that $1 + 2 + ... + k = \frac{k(k+1)}{2}$, so $1 + 2 + ... + k + (k + 1) = \frac{k(k+1)}{2} + (k + 1) = \frac{k^2 + 3k + 2}{2} = \frac{(k+1)(k+2)}{2}$. This establishes the result for all natural numbers n by induction. □

Theorem 2.3. *Let $m, n \in \mathbb{N}$.*

(a) $m + n \in \mathbb{N}$

(b) $mn \in \mathbb{N}$.

(c) If $n > m$ then $n = m + k$ for some $k \in \mathbb{N}$. In other words, $n - m \in \mathbb{N}$.

Proof. Fix $m \in \mathbb{N}$.

(a) Let $P(n)$ be the statement that $m+n \in \mathbb{N}$. We know $m+1 \in \mathbb{N}$ since $\mathbb{N}$ is inductive. Assume that $m+k \in \mathbb{N}$ for some $k \in \mathbb{N}$. Then $m+k+1 \in \mathbb{N}$ because $\mathbb{N}$ is inductive. Hence, $m + n \in \mathbb{N}$ for all $n \in \mathbb{N}$. Since the choice of m was arbitrary, $m + n \in \mathbb{N}$ for all $m, n \in \mathbb{N}$

(b) Let $Q(n)$ be the statement that $mn \in \mathbb{N}$. We know that $m(1) \in \mathbb{N}$. Assume that $mk \in \mathbb{N}$. Then $m(k+1) = mk + m \in \mathbb{N}$ since we know that mk and m are natural numbers and we have shown that the sum of natural numbers is a natural number. It follows that $mn \in \mathbb{N}$ for all $n \in \mathbb{N}$. Since the choice of m was arbitrary, it follows that $mn \in \mathbb{N}$ for all $m, n \in \mathbb{N}$

(c) Let $S = \{n \in \mathbb{N} | n > m \text{ and } n - m \notin \mathbb{N}\}$. Suppose that $S \neq \emptyset$. Then S has a least element t since $\mathbb{N}$ is well-ordered. We know that t is not $m+1$ since $m+1-m = 1 \in \mathbb{N}$, so t is at least $m+2$. Hence, $t-1 > m$ and, since $t-1 \notin S$, we know that $(t-1)-m \in \mathbb{N}$, from which it follows that $t-1-m+1 = t-m \in \mathbb{N}$, contradicting the assumption that $t \in S$. The result follows.

□

Definition 13

The *integers* $\mathbb{Z}$ are the set $\mathbb{N} \cup -\mathbb{N} \cup \{0\} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}$. The *rational numbers* are the set $\mathbb{Q} = \{\frac{p}{q} \in \mathbb{R} | p \in \mathbb{Z} \text{ and } q \in \mathbb{N}\}$.

Theorem 2.4. *(a) For any integer k, there are no integers between k and $k+1$ or between k and $k-1$.*

(b) If m, n are integers then $m+n$ and mn are integers.

(c) Let $k \in \mathbb{Z}$. Then if j is an integer so that $j > k$ then $j = k + m$ for some natural number m.

Proof. (a) The only positive integers are natural numbers and for every negative integer k we know that $-k \in \mathbb{N}$ by definition. Thus, 1 is the least positive integer and -1 is the greatest negative integer, so there are no integers between -1 and 0 or between 0 and 1. If $k \in \mathbb{N}$ then by Theorem 7.3, it follows that there are no integers between k and $k+1$ or between k and $k-1$. If $-k \in \mathbb{N}$ then again by Theorem 7.3 it follows that there are no integers between $-k$ and $-(k+1)$ and no integers between $-k$ and $-(k-1)$, which means that there are no integers between k and $k+1$ or between k and $k-1$.

(b) If $m = 0$ then $nm = 0$ and $m + n = n$. If $m, n \in \mathbb{N}$ then the result follows by Theorem 2.3. If $m, n \in -\mathbb{N}$ then $mn = (-m)(-n) \in \mathbb{N}$ by the same theorem, and $m+n = -(-m+-n)$. Since $-m, -n \in \mathbb{N}$ we know that $-m-n \in \mathbb{N}$, so $m+n \in -\mathbb{N} \subset \mathbb{Z}$. If $m \in \mathbb{N}$ and $n \in -\mathbb{N}$ then $m(-n) \in \mathbb{N}$, so $mn \in -\mathbb{N}$. Also, $m+n = m-(-n) \in \mathbb{N}$ if $m > -n$ by the third part of Theorem 2.3. If $m = -n$ then $m+n = 0 \in \mathbb{Z}$. If $m < -n$ then $-n-m \in \mathbb{N}$ which means that $m+n \in -\mathbb{N} \subset \mathbb{Z}$.

(c) By (b) we know that $j-k \in \mathbb{Z}$ and since $j > k$ we know that $j-k > 0$ which means that $j-k = m \in \mathbb{N}$ since the only positive integers are natural numbers.

□

Theorem 2.5. *Generalized and Strong Induction. Let $j \in \mathbb{Z}$. For each $n \in \mathbb{Z}$, let $P(n)$ be a statement so that (a) $P(j)$ is true, and one of the following is true:*

(b) whenever $P(k)$ is true for an integer $k \geq j$, it follows that $P(k+1)$ is also true, or

(b)' whenever $P(i)$ is true for all integers i such that $j \leq i \leq k$, it follows that $P(k+1)$ is also true.

Then it follows that $P(n)$ is true for all integers $n \geq j$.

Proof. First, assume (a) and $(b)'$ are true. Let $Q(n)$ be the statement that the statements $P(j), P(j+1), ..., P(j+n-1)$ are true. Then $Q(1)$ is that $P(j)$ is true, which follows from (a). Assuming that $Q(k)$ is true we know that $P(j), P(j+1), ..., P(j+k-1)$ is true, so $P(j+k)$ is also true by (b)', which means that $Q(k+1)$ is true. Hence, $Q(n)$ is true for all natural numbers n by induction, and so $P(i)$ is true for all integers $i \geq j$ by Theorem 2.4 part (c).

Assume (a) and (b). Since (b) implies $(b)'$ the result follows from the preceding argument. □

"Generalized" induction is where we start the induction at an integer other than one (we show (a) and (b)). "Strong" induction is where we assume the result has been shown for all $j \leq n \leq k$ instead of just for $n = k$ in order to prove the result is true when $n = k+1$ (we show (a) and (b)'). Here is an example using strong induction. The proof could be done just using generalized induction in this case, but we will use strong induction to illustrate how it can be used.

Example 2.1. *Let $n \in \mathbb{N}$ and let $n \geq 8$. Then prove that $n = 3i+5j$ for some non-negative integers i and j.*

Solution. If $n = 8$ we see that $3(1)+5(1) = 8$, so the theorem is true. Also, $3(3)+5(0) = 9$, and $3(0)+2(5) = 10$. Assume the result is true for all natural numbers m so that $8 \leq m \leq k$, where $k \geq 10$. Then we know that there are non-negative integers i, j so that $k-2 = 3i+5j$ since $k-2 \geq 8$, and thus $k+1 = 3(i+1)+5j$. The result follows by strong induction. □

Definition 14

For any $x \in \mathbb{R}$ we define $x^1 = x$ and for each $n \in \mathbb{N}$, if $n > 1$ and x^{n-1} is known then the nth power x^n of x is defined to be to be $x(x^{n-1})$. If $x \neq 0$ then we define $x^0 = 1$ and $x^{-n} = \dfrac{1}{x^n}$ for each $n \in \mathbb{N}$. If $x \geq 0$ we define an nth root $x^{\frac{1}{n}}$ of x to be a positive real number whose nth power is x. We use the notation $x^{\frac{p}{q}} = (x^{\frac{1}{q}})^p$ for $q \in \mathbb{N}$ and $p \in \mathbb{Z}$ when these exist.

Uniqueness and existence of nth roots are established in exercises.

Though 0^0 is undefined, it is a common convention (which by default we use unless otherwise specified in this text) to use $g(x) = (f(x))^0$ to denote the function $g(x) = 1$ (so we define the function $g(x)$ to be one even when $f(x) = 0$ in this notation). This makes many theorem statements less messy, though it is somewhat confusing because it means that, most of the time, when we write x^0 what we mean is x^0 if $x \neq 0$ and 1 if $x = 0$.

Note that we have defined positive integer powers of x recursively, which is a notion described in more detail in the Induction portion of the Supplementary Materials section.

Theorem 2.6. *$\mathbb{N}$ is not bounded above.*

Proof. Suppose $\mathbb{N}$ is bounded above. Then $\mathbb{N}$ has a least upper bound u. Hence, there is some $n \in \mathbb{N}$ so that $n > u - 1$ by the Approximation Property, which means that $n + 1 > u$, which is a contradiction to u being a upper bound for $\mathbb{N}$ since $n + 1 \in \mathbb{N}$.

□

Induction can be used to prove many interesting results. One example is the Binomial Theorem, which is our next objective.

Definition 15

The notation $\binom{n}{k}$ means $\dfrac{n!}{(n-k)!k!}$ if $k \leq n$ and n and k are non-negative integers (where $n! = n(n-1)(n-2)...(1)$ if $n \geq 1$ and $0! = 1$).

Theorem 2.7. *For any $n, k \in \mathbb{N}$ so that $n \geq k$ it is true that* $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$

Proof. $\dfrac{n!}{(n-k)!k!} + \dfrac{n!}{(n-k+1)!(k-1)!} = \dfrac{n!(n-k+1+k)}{(n-k+1)(n-k)!k(k-1)!} = \dfrac{(n+1)!}{(n+1-k)!k!}$.

□

Theorem 2.8. *The Binomial Theorem. For every positive integer n,* $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i$.

Proof. We proceed by induction on n. If $n = 1$ we note that $(x + y) = \binom{1}{0} x^1 y^0 + \binom{1}{1} x^0 y^1$ is true. Assume that $(x + y)^k = \sum_{i=0}^{k} \binom{k}{i} x^{k-i} y^i$. Multiplying by $(x + y)$ on both sides of this equation and rearranging terms yields $\binom{k+1}{0} x^{k+1} y^0 + \binom{k+1}{k+1} x^0 y^{k+1} + \sum_{i=1}^{k} (\binom{k}{i-1} + \binom{k}{i}) x^{k+1-i} y^i$ on the right side of the equation, which, by Theorem 2.7, is equal to $\sum_{i=0}^{k+1} \binom{k+1}{i} x^{k+1-i} y^i$. The result follows by induction.

□

One useful consequence of the fact that the natural numbers are not bounded above is that the rational numbers are dense in the real numbers, meaning that there is a rational number in each non-empty open interval. We will present two proofs.

Theorem 2.9. *Let a, b be real numbers so that $a < b$. Then there is a rational number between a and b.*

Proof. If $a < 0 < b$ then since 0 is rational, we are finished. Next, assume that $0 \leq a < b$. Since $\mathbb{N}$ is not bounded above by Theorem 2.6, we can find a positive integer $q > \frac{1}{b-a}$, so that $\frac{1}{q} < b - a$. Also since $\mathbb{N}$ is not bounded above, we can find a positive integer $m > qa$ so that $\frac{m}{q} > a$. Let $S = \{i \in \mathbb{N} | \frac{i}{q} > a\}$. Since S is a non-empty subset of $\mathbb{N}$ it follows that S has a least element p. Note that if $p = 1$ then $\frac{p-1}{q} = 0 \leq a$, and otherwise $p - 1 \in \mathbb{N}$, so $\frac{p-1}{q} \leq a$ since p is the least element of S. Hence, $\frac{p-1}{q} + \frac{1}{q} < a + b - a = b$, so $a < \frac{p}{q} < b$. Finally, assume that $a < b \leq 0$. Then by the preceding argument we can find a rational number r so that $-b < r < -a$, so $a < -r < b$. Hence, in each case the result follows.

□

Alternate proof:

Proof. Since $\mathbb{N}$ is not bounded above, we can find $q \in \mathbb{N}$ so that $q > \frac{1}{b-a}$, so $qb - qa > 1$. By Exercise 2.6, we know that there is a first integer m so that $m \geq qb$. Hence, $m - 1$ is an integer so that $m - 1 < qb$. Since $m - 1 \geq qb - 1 > qa$ it follows that $qa < m - 1 < qb$, and therefore $a < \frac{m-1}{q} < b$. □

We sometimes want to talk about divisibility and prime numbers, so we will define those terms here. Some exercises that are good induction examples use divisibility.

Definition 16

We say that an integer m is *divisible* by an integer k if $\frac{m}{k}$ is an integer (also written as $m = qk$ for some integer q), and we refer to k as a *divisor* of the integer m and say that k *divides* m or m is divisible by k. We use the notation $k|m$ to denote "k divides m."

We also use some of the same terms for any ratio $\frac{x}{y}$ of real numbers x and y and refer to x as the numerator or dividend and y as the denominator or divisor, but when we are referring to a divisor of an integer we normally mean the definition above (an integer which divides into the numerator evenly).

Definition 17

We say that a natural number $p > 1$ is *prime* if the only natural numbers that divide p are 1 and p. A positive integer is *composite* if it is not prime (meaning it is the product of two natural numbers, neither of which are one). We also refer to any non-zero integer as being composite if its absolute value is composite. The *greatest common divisor* of non-zero integers a, b is the largest natural number that is a divisor of both a and b. We say an integer is *even* if it is divisible by two, and *odd* if it is not.

We will not be spending a lot of time on divisibility, but the Fundamental Theorem of Arithmetic is a good theorem using strong induction for its proof and is helpful in advanced calculus. So, a proof of this result is also found in the Supplementary Materials section for the single variable development.

Exercises:

Exercise 2.1. *For each natural number* n, $1^2 + 2^2 + 3^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$.

Exercise 2.2. *For each natural number* n, $1^3 + 2^3 + 3^3 + ... + n^3 = \frac{n^2(n+1)^2}{4}$.

Exercise 2.3. $5^n - 1$ *is divisible by* 4 *for all natural numbers* n.

Exercise 2.4. *Let* $n \in \mathbb{Z}$. *Then* n *odd if and only if* $n+1$ *is even and* n *is even if and only if* $n+1$ *is odd.*

Exercise 2.5. *Define the Fibonacci sequence to be the function* $f : \mathbb{N} \to \mathbb{R}$ *by the following recursive definition. We use the notation* $f(n) = a_n$ *and define* $a_1 = a_2 = 1$ *and define* $a_n = a_{n-1} + a_{n-2}$ *for integers* $n > 2$. *Prove that* $\frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n] = a_n$ *for every positive integer* n. *(Assume, for now, that we know there is a positive number whose square is five - a fact that will be proven later).*

Exercise 2.6. *Let* S *be a non-empty set of integers which is bounded below. Then* S *has a first element.*

Exercise 2.7. *Let* $m, n \in \mathbb{Z}$. *If* m *is even then* mn *is even. If* m *and* n *are both odd then* mn *is odd.*

Exercise 2.8. *Let* $m \in \mathbb{Z}$. *Then, for every positive integer* n, *it is true that* m *is even if and only if* m^n *is even.*

Exercise 2.9. *Let* S *be a non-empty set of integers which is bounded above. Then* S *has a last element.*

Exercise 2.10. *For any natural number* n *it is true that* $n^3 + 3n$ *is even.*

Exercise 2.11. *For any natural number* n *it is true that* $n^3 + 2n$ *is divisible by three.*

Exercise 2.12. *Show* $n! > 2^n$ *for any natural number* $n \geq 4$.

Exercise 2.13. *Let* $0 < a < b$. *Then, for every positive integer* n, *show that* $0 < a^n < b^n$. *If there are positive numbers* $a^{\frac{1}{n}}$ *and* $b^{\frac{1}{n}}$ *whose* n*th powers are* a *and* b *respectively, then* $0 < a^{\frac{1}{n}} < b^{\frac{1}{n}}$.

Hints:

Hint to Exercise 2.1. *For each* $n \in \mathbb{N}$ *it is true that* $1^2+2^2+3^2+...+n^2 = \dfrac{n(n+1)(2n+1)}{6}$.

Use induction.

Hint to Exercise 2.2. *For each natural number* n, $1^3 + 2^3 + 3^3 + ... + n^3 = \dfrac{n^2(n+1)^2}{4}$.

Use induction.

Hint to Exercise 2.3. $5^n - 1$ *is divisible by* 4 *for all natural numbers* n.

Use induction. Remember that for a natural number to be divisible by 4 is the same as that number being equal to $4m$ for some integer m.

Hint to Exercise 2.4. *Let* $n \in \mathbb{Z}$. *Then* n *odd if and only if* $n+1$ *is even and* n *is even if and only if* $n+1$ *is odd.*

Divide an even number plus one by two. Explain why an integer plus one half cannot be an integer.

Hint to Exercise 2.5. *Define the Fibonacci sequence to be the function* $f : \mathbb{N} \to \mathbb{R}$ *by the following recursive definition. We use the notation* $f(n) = a_n$ *and define* $a_1 = a_2 = 1$ *and define* $a_n = a_{n-1} + a_{n-2}$ *for integers* $n > 2$. *Prove that* $\frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n] = a_n$ *for every positive integer* n. *(Assume, for now, that we know there is a positive number whose square is five - a fact that will be proven later).*

Use strong induction.

Hint to Exercise 2.6. *Let* S *be a non-empty set of integers which is bounded below. Then* S *has a first element.*

One approach would be to look at $S + k$, where k is a large enough natural number so that all elements of $S + k$ are positive (once you have explained why there is a natural number k so that all elements of $S + k$ are positive).

Hint to Exercise 2.7. *Let* $m, n \in \mathbb{Z}$. *If* m *is even then* mn *is even. If* m *and* n *are both odd then* mn *is odd.*

Use the definition of being even and odd and Exercise 2.4.

Hint to Exercise 2.8. *Let* $m \in \mathbb{Z}$. *Then, for every positive integer* n, *it is true that* m *is even if and only if* m^n *is even.*

Use Exercise 2.7 and induction.

Hint to Exercise 2.9. *Let S be a non-empty set of integers which is bounded above. Then S has a last element.*

Consider the set of integers which are upper bounds for S, and use well ordering.

Hint to Exercise 2.10. *For any natural number n it is true that $n^3 + 3n$ is even.*

Use induction or prove that the sum of odd integers is even.

Hint to Exercise 2.11. *For any natural number n it is true that $n^3 + 2n$ is divisible by three.*

Use induction to show that $n^3 + 2n$ is always three times a natural number.

Hint to Exercise 2.12. *Show $n! > 2^n$ for any natural number $n \geq 4$.*

Use generalized induction.

Hint to Exercise 2.13. *Let $0 < a < b$. Then, for every positive integer n, show that $0 < a^n < b^n$. If there are positive numbers $a^{\frac{1}{n}}$ and $b^{\frac{1}{n}}$ whose nth powers are a and b respectively, then $0 < a^{\frac{1}{n}} < b^{\frac{1}{n}}$.*

Use induction to show $0 < a^n < b^n$. Then use that result to prove $0 < a^{\frac{1}{n}} < b^{\frac{1}{n}}$ (try contradiction to see the connection).

Solutions:

Solution to Exercise 2.1. *For each* $n \in \mathbb{N}$ *it is true that* $1^2 + 2^2 + 3^2 + ... + n^2 = \dfrac{n(n+1)(2n+1)}{6}$.

Proof. Proceed by induction. If $n = 1$ then $1^2 = \dfrac{(1)(2)(3)}{6}$. Assume that $1^2 + 2^2 + 3^2 + ... + k^2 = \dfrac{k(k+1)(2k+1)}{6}$ for some $k \in \mathbb{N}$. Then $1^2 + 2^2 + 3^2 + ... + k^2 + (k+1)^2 = \dfrac{k(k+1)(2k+1)}{6} + k^2 + 2k + 1 = \dfrac{2k^3 + 3k^2 + k + 6k^2 + 12k + 6}{6} = \dfrac{2k^3 + 9k^2 + 13k + 6}{6} = \dfrac{(k+1)(k+2)(2(k+1)+1)}{6}$. The result follows by induction. □

Solution to Exercise 2.2. *For each natural number* n, $1^3 + 2^3 + 3^3 + ... + n^3 = \dfrac{n^2(n+1)^2}{4}$.

Proof. If $n = 1$ we have $1^3 = \dfrac{1^2(1+1)^2}{4}$. Assume that for a positive integer k it is true that $1^3 + 2^3 + 3^3 + ... + k^3 = \dfrac{k^2(k+1)^2}{4}$. Then $1^3 + 2^3 + 3^3 + ... + k^3 + (k+1)^3 = \dfrac{k^2(k+1)^2}{4} + (k+1)^3 = \dfrac{k^4 + 2k^3 + k^2}{4} + k^3 + 3k^2 + 3k + 1 = \dfrac{k^4 + 2k^3 + k^2 + 4k^3 + 12k^2 + 12k + 4}{4} = \dfrac{k^4 + 6k^3 + 13k^2 + 12k + 4}{4} = \dfrac{(k+1)^2(k+2)^2}{4}$. The result follows by induction. □

Solution to Exercise 2.3. $5^n - 1$ *is divisible by* 4 *for all natural numbers* n.

Proof. Proceeding by induction, when $n = 1$ we have $5^1 - 1 = 4$ is divisible by four since $\dfrac{4}{4} = 1$. Assume that for some $n = k \in \mathbb{N}$ we know that $5^k - 1 = 4m$ for some natural number m. Then it follows that $5(5^k - 1) = 20m$, so $5^{k+1} - 1 = 20m + 4 = 4(5m + 1)$. Thus, $5^{k+1} - 1$ is divisible by four, so the statement holds for all natural numbers n by induction. □

Solution to Exercise 2.4. *Let* $n \in \mathbb{Z}$. *Then* n *odd if and only if* $n + 1$ *is even and* n *is even if and only if* $n + 1$ *is odd.*

Proof. We proceed by induction for natural numbers n. Note that $n = 1$ is not even since $\dfrac{1}{2} < 1$ which means it cannot be a positive integer (and $\dfrac{1}{2}$ is positive, so it is not an integer), whereas $1 + 1$ is even since $\dfrac{2}{2} = 1$ and $2 + 1$ is odd since $\dfrac{3}{2} = 1 + \dfrac{1}{2}$ which is between one and two and so is not an integer. Using strong induction we assume that the statement has been established for all $n \leq k \in \mathbb{N}$ where $k \geq 2$. If $k + 1$ is even then $\dfrac{k+2}{2} = \dfrac{k+1}{2} + \dfrac{1}{2}$ where $\dfrac{k+1}{2} \in \mathbb{N}$. However, since $\dfrac{1}{2} < 1$ we know that $\dfrac{k+1}{2} + \dfrac{1}{2} < \dfrac{k+1}{2} + 1$ and we also know there are no integers between $\frac{k+1}{2}$ and $\frac{k+1}{2}+1$, so $\frac{k+2}{2}$ is not an integer, which means $k+2$ is odd. If $k+1$ is odd then from the induction hypothesis we know that k is even, which means that $\frac{k+2}{2} = \frac{k}{2}+1$ is the sum of two natural numbers which is a natural number, so $k+2$ is even.

Having shown this is true for natural numbers, we first note that for any negative integer n, it is true that if $-n = 2m$ then $n = -2m$, so n is even if and only if $-n$ is even. If $m = 0$ then m is even and $m+1$ is odd, and if $m = -1$ then m is odd and $m+1 = 0$ is even. If m is an integer less than -1 then $-m, (-m+1), (-m-1) \in \mathbb{N}$ and so we know that $-m$ is even if and only if $-m+1$ and $-m-1$ are odd. We also know that m is even if and only if $-m$ is even, and $-m+1$ and $-m-1$ are odd if and only if $m-1$ and $m+1$ are odd. Thus, m is even if and only if $m+1$ and $m-1$ are odd. □

Solution to Exercise 2.5. *Define the Fibonacci sequence to be the function $f : \mathbb{N} \to \mathbb{R}$ by the following recursive definition. We use the notation $f(n) = a_n$ and define $a_1 = a_2 = 1$ and define $a_n = a_{n-1}+a_{n-2}$ for integers $n > 2$. Prove that $\frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n] = a_n$ for every positive integer n. (Assume, for now, that we know there is a positive number whose square is five - a fact that will be proven later).*

Proof. We proceed by strong induction to show $\frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^j - (\frac{1-\sqrt{5}}{2})^j] = a_j$ for every positive integer $j \leq n$. The statement is immediate if $n = 1$ or $n = 2$. Assume the statement is true for all $1 \leq j \leq k$ for some $k \geq 2$. Then it follows that $a_{k+1} = a_k + a_{k-1} = \frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^k - (\frac{1-\sqrt{5}}{2})^k + (\frac{1+\sqrt{5}}{2})^{k-1} - (\frac{1-\sqrt{5}}{2})^{k-1}] = \frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^{k-1}(\frac{1+\sqrt{5}}{2}+1) - (\frac{1-\sqrt{5}}{2})^{k-1}(\frac{1-\sqrt{5}}{2}+1)]$. But $(\frac{(1+\sqrt{5}}{2})^2 = \frac{1+5+2\sqrt{5}}{4} = \frac{3+\sqrt{5}}{2} = \frac{1+\sqrt{5}}{2}+1$ and $(\frac{(1-\sqrt{5}}{2})^2 = \frac{1+5-2\sqrt{5}}{4} = \frac{3-\sqrt{5}}{2} = \frac{1-\sqrt{5}}{2}+1$. Thus, $a_{k+1} = \frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^{k-1}(\frac{1+\sqrt{5}}{2}+1) - (\frac{1-\sqrt{5}}{2})^{k-1}(\frac{1-\sqrt{5}}{2}+1)] = \frac{1}{\sqrt{5}}[(\frac{1+\sqrt{5}}{2})^{k+1} - (\frac{1-\sqrt{5}}{2})^{k+1}]$ as desired. The result follows. □

Solution to Exercise 2.6. *Let S be a non-empty set of integers which is bounded below. Then S has a first element.*

Proof. Let m be the greatest lower bound of S. By the Approximation Property there is some integer $n \in S$ so that $m \leq n < m+1$. Suppose $n > m$. Then $n-1 < m < n$ and there are no integers in $(n-1, n)$ which means that S contains no points less than m or in $(n-1, n)$ and hence no points less than n. Thus, n is a lower bound for S, contradicting the fact that m is the greatest lower bound of S, We conclude that $n = m$ and therefore $m \in S$.

□

Solution to Exercise 2.7. *Let $m, n \in \mathbb{Z}$. If m is even then mn is even. If m and n are both odd then mn is odd.*

Proof. If j, k are integers and j is even then $j = 2s$ and so $jk = 2sk$ which means that jk is even.

If j, k are odd integers then $j = 2s + 1$ and $k = 2t + 1$ for some integers s, t by Exercise 2.4, which means that $jk = 4st + 2(s + t) + 1$, which is odd since $4st + 2(s + t) = 2(2st + s + t)$ is even (by Exercise 2.4). □

Solution to Exercise 2.8. *Let $m \in \mathbb{Z}$. Then, for every positive integer n, it is true that m is even if and only if m^n is even.*

Proof. If m is even then m^1 is even. Assuming m^k is even for some natural number k, we have $m^{k+1} = m^k m = m^{k+1}$ is a product of two even integers and is even by Exercise 2.7. If m is odd then $m^1 = m$ is odd. Assuming m^k is odd for some natural number k we have that $m^k m = m^{k+1}$ is a product of odd integers and is odd by Exercise 2.7. Thus, by induction it follows that m^n is even for all $n \in \mathbb{N}$ if m is even, and m^n is odd for all $n \in \mathbb{N}$ if m is odd. □

Solution to Exercise 2.9. *Let S be a non-empty set of integers which is bounded above. Then S has a last element.*

Proof. Let $u = \sup(S)$. By the approximation property we can find an element $s \in S$ so that $u - 1 < s \leq u$. But there are no integers in $(s, s + 1)$, which means that every element of S is less than or equal to s, and therefore $s = u$, which is the largest element of S. □

Solution to Exercise 2.10. *For any natural number n it is true that $n^3 + 3n$ is even.*

Proof. Proceeding by induction, $1^3 + 3(1) = 4 = 2(2)$ is even. Assume that $k^3 + 3k = 2m$ for some natural number m. Then $(k + 1)^3 + 3(k + 1) = k^3 + 3k^2 + 3k + 1 + 3k + 3$ $= k^3 + 3k + (3k^2 + 3k + 4) = k^3 + 3k + 3k(k + 1) + 4 = 2m + 3k(k + 1) + 4$. By Exercise 2.4, we know that either k or $k + 1$ is even. Thus, there is some natural number j so that either $k = 2j$ or $k + 1 = 2j$. Hence, either $(k + 1)^3 + 3(k + 1) = 2(m + 3j(k + 1) + 2)$ or $(k + 1)^3 + 3(k + 1) = 2(m + 3jk + 2)$, which means that $(k + 1)^3 + 3(k + 1)$ is even. The result follows by induction. □

Solution to Exercise 2.11. *For any natural number n it is true that $n^3 + 2n$ is divisible by three.*

Proof. We proceed by induction. We know that $1^3 + 2(1) = 3$ is divisible by 3. Assume that $k^3 + 2k = 3m$ for some natural number m. Then $(k+1)^3 + 2(k+1) = k^3 + 3k^2 + 3k + 1 + 2k + 2$ $= (k^3 + 2k) + 3(k^2 + k + 1) = 3(m + k^2 + k + 1)$, which is divisible by 3. The result follows by induction. □

Solution to Exercise 2.12. *Show $n! > 2^n$ for any natural number $n \geq 4$.*

Proof. We proceed by generalized induction. We know $4! > 2^4$ is true. Assume the statement $k! > 2^k$ is true for an integer $k \geq 4$. Then $(k+1)! = (k+1)(k!) \geq 5(k!) > 2(k!) > 2(2^k) = 2^{k+1}$ by the induction hypothesis. Thus, $n! > 2^n$ for any natural number $n \geq 4$ by generalized induction. □

Solution to Exercise 2.13. *Let $0 < a < b$. Then, for every positive integer n, show that $0 < a^n < b^n$. If there are positive numbers $a^{\frac{1}{n}}$ and $b^{\frac{1}{n}}$ whose nth powers are a and b respectively, then $0 < a^{\frac{1}{n}} < b^{\frac{1}{n}}$.*

Proof. We proceed to show $0 < a^n < b^n$ by induction. If $n = 1$ then $0 < a < b$ is given to be true. Assume this is true when $n = k$. Then $0 < a^k < b^k$. Since $a > 0$ we have $(0)(a) < a^{k+1}$. Since $a < b$ we know that $a(a^k) < b(a^k)$. Since $a^k < b^k$ we know that $b(a^k) < b(b^k)$. Combining these inequalities gives $0 < a^{k+1} < b^{k+1}$. By induction the result holds for all positive integers n.

We are given that these nth roots are positive. By the preceding part of the argument, if $b^{\frac{1}{n}} \leq a^{\frac{1}{n}}$ then $(b^{\frac{1}{n}})^n \leq (a^{\frac{1}{n}})^n$ so $b \leq a$, a contradiction. Hence, it must follow that $0 < a^{\frac{1}{n}} < b^{\frac{1}{n}}$. □

Chapter 3

Sequences

Definition 18

A *sequence* is a function whose domain is $\mathbb{N}$. If $f : \mathbb{N} \to \mathbb{R}$ is a sequence then we use the notation $\{x_n\}$ to denote the function f, where $f(n) = x_n$. We refer to n as the *index* of the sequence member x_n. Depending on context we may also use $\{x_n\}$ to refer to the range of f. If $g : \mathbb{N} \to \mathbb{N}$ is an increasing function then we say that the sequence $\{x_{g(n)}\}$ is a *subsequence* of $\{x_n\}$, and normally write $g(i) = n_i$. Essentially, in a subsequence, infinitely many terms from the original sequence are listed in the same relative order.

We say that $\{x_n\}$ *converges* to c, often written $\{x_n\} \to c$ (also written $\lim\limits_{n\to\infty} x_n = c$) if for every $\epsilon > 0$ there is some $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - c| < \epsilon$.

We say that a statement $P(x)$ about x is true for $x \in D$ *sufficiently large* (or if $x \in D$ is sufficiently large) if there is a number M so that if $x > M$ and $x \in D$ then $P(x)$ is true. We say that a statement $P(x)$ is true for $x \in D$ *sufficiently small* if there is a number δ so that if $x < \delta$ then $P(x)$ is true. We say statement $P(x)$ is true if $x \in D$ is sufficiently close to c if there is a $\delta > 0$ so that $P(x)$ is true for all $x \in D$ so that $|x - c| < M$. If no set D is specified it is understood that $D = \mathbb{R}$.

If D is understood we may not mention what D is. For instance, when talking about sequence indices, those indices must be real numbers. For example, rather than say $x_n > 2$ for $n \in \mathbb{N}$ sufficiently large we might just say $x_n > 2$ for n sufficiently large (since it is known that there is no such thing as x_n for $n \notin \mathbb{N}$). Another way of stating the definition of convergence that sounds more intuitive for some people is thus:

We say $\{x_n\} \to c$ if for every positive distance ϵ from c, every sequence member x_n is distance less than ϵ from c if its index n is sufficiently large.

In the theorems that follow, sequences listed are assumed to be sequences of real numbers unless otherwise stated. Likewise, if we indicate that a sequence converges to something then it is understood the thing converged to is a real number.

When establishing sequence convergence for a sequence $\{x_n\}$ to a point p, one normally starts by declaring an arbitrary $\epsilon > 0$ and then shows the existence of a corresponding $k \in \mathbb{N}$ so that sequence members of index higher than k (sequence members listed later in the sequence order than the kth term) are within a distance ϵ of the point p. The integer k is different for different ϵ values, so k can be thought of as a function of ϵ. When first encountering sequence convergence, people sometimes think that they just have to find a k so that x_k is within some particular distance of p, but that isn't what is needed. It has to be shown that no matter what distance ϵ from p we start with, there is always some corresponding $k = k(\epsilon)$ so that all x_n sequence members with $n \geq k$ are within distance ϵ of the point p. We are really showing the existence of a function $k(\epsilon) : (0, \infty) \to \mathbb{N}$ in this manner because every $\epsilon > 0$ must have an associated k. For brevity in notation, we don't normally write $k(\epsilon)$ to refer to k as a function of ϵ and instead simply show the existence of a k corresponding to an arbitrary $\epsilon > 0$ satisfying the definition of convergence.

Theorem 3.1. $\{\frac{1}{n}\} \to 0$.

Proof. Let $\epsilon > 0$. Since $\mathbb{N}$ is not bounded above, we can find $k \in \mathbb{N}$ so that $k > \frac{1}{\epsilon}$, so $\frac{1}{k} < \epsilon$. Hence, if $n \geq k$ then $0 < \frac{1}{n} \leq \frac{1}{k} < \epsilon$, so $|\frac{1}{n} - 0| < \epsilon$. □

Theorem 3.2. *Let $j \in \mathbb{N}$. If $x_n = c$ for each $n \in \mathbb{N}$ so that $n \geq j$ then $\{x_n\} \to c$.*

Proof. Let $\epsilon > 0$. Since $|x_n - c| = |c - c| = 0 < \epsilon$ for all $n \geq j$, it follows that $\{x_n\} \to c$. □

Theorem 3.3. *The sequence $\{x_n\} \to c$ if and only if $\{x_n - c\} \to 0$.*

Proof. We know that $\{x_n\} \to c$ if and only if for every $\epsilon > 0$ there is some $N \in \mathbb{N}$ so that if $n \geq N$ then $|x_n - c| < \epsilon$, which is true if and only if for every $\epsilon > 0$ there is some $N \in \mathbb{N}$ so that if $n \geq N$ then $|(x_n - c) - 0| < \epsilon$, which is true if and only if $\{x_n - c\} \to 0$. □

Theorem 3.4. *If $\{x_n\}$ is bounded and $\{y_n\} \to 0$ then $\{x_n y_n\} \to 0$.*

Proof. Choose $M > 0$ so that $-M \leq x_n \leq M$ for all $n \in \mathbb{N}$. Let $\epsilon > 0$. Choose $N \in \mathbb{N}$ so that if $n \geq N$ then $|y_n - 0| < \frac{\epsilon}{M}$. Then $|x_n y_n - 0| < M\frac{\epsilon}{M} = \epsilon$, so $\{x_n y_n\} \to 0$. □

The proof of the next theorem uses the fact that a finite set always has a first and a last point, which is not really addressed until the end of this chapter in Theorem 3.30. Though we normally prefer to put theorems using a result after the result has been proven, in this case we put the proof later with other proofs about cardinality in order to avoid taking the direction of the subject matter on a tangent in the middle of developing sequences. However, the proof of Theorem 3.30 is self-contained (so the argument is not circular) and nothing is lost by reading that proof first if preferred. Most of the time we do not quote Theorem 3.30 explicitly (it is normal in texts to assume that this is understood).

Theorem 3.5. *If $\{x_n\} \to p$ then $\{x_n\}$ is bounded.*

Proof. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - p| < 1$, so $p - 1 < x_n < p + 1$. Let $M = \max(x_1, x_2, ..., x_k, p + 1)$ and $m = \min(x_1, x_2, ..., x_k, p - 1)$. Then if $n \leq k$ we know that $m \leq \min(x_1, x_2, ..., x_k) \leq x_n \leq \max(x_1, x_2, ..., x_k) \leq M$ and if $n > k$ then $m \leq p - 1 < x_n < p + 1 \leq M$. Hence, for all $n \in \mathbb{N}$ it follows that $m \leq x_n \leq M$.

□

Theorem 3.6. *The Squeeze Theorem. Let $\{x_n\} \to c$ and $\{z_n\} \to c$, and let $\{y_n\}$ be a sequence so that for some positive integer j, if $n \geq j$ then $x_n \leq y_n \leq z_n$ or $z_n \leq y_n \leq x_n$. Then $\{y_n\} \to c$.*

Proof. Let $\epsilon > 0$. Choose $k_1 \in \mathbb{N}$ so that if $n \geq k_1$ then $|x_n - c| < \epsilon$, and choose $k_2 \in \mathbb{N}$ so that if $n \geq k_2$ then $|z_n - c| < \epsilon$. If $n \geq k = \max\{k_1, k_2, j\}$ then $c - \epsilon < x_n \leq y_n \leq z_n < c + \epsilon$ or $c - \epsilon < z_n \leq y_n \leq x_n < c + \epsilon$, so $|y_n - c| < \epsilon$.

□

The Squeeze Theorem can be used to prove the convergence of many sequences and later we can use it to prove limits of functions which are not sequences. Note that in many texts a slightly weaker theorem is referred to as the Squeeze Theorem, but the version we have proven is useful in more instances.

Example 3.1. *Prove $\{\frac{1}{2n^2+1}\} \to 0$.*

Solution. First, note that since $n \geq 1$ and $0 < 1 < 2$ we have $n \leq n^2 < 2n^2 < 2n^2 + 1$, so $0 < \frac{1}{2n^2+1} < \frac{1}{n}$. Since we have shown that $\{\frac{1}{n}\} \to 0$ and $\{0\} \to 0$ it follows that $\{\frac{1}{2n^2+1}\} \to 0$ by the Squeeze Theorem. □

Theorem 3.7. *The Comparison Theorem. Let $\{x_n\} \to c$ and $\{y_n\} \to d$, so that, for some $j \in \mathbb{N}$, $x_n \leq y_n$ for all $n \geq j$. Then $c \leq d$.*

Proof. Suppose $d < c$. We can find $k_1 \in \mathbb{N}$ so that if $n \geq k_1$ then $|x_n - c| < \frac{c-d}{2}$, meaning that $c - \frac{c-d}{2} < x_n < c + \frac{c-d}{2}$. We can also find $k_2 \in \mathbb{N}$ so that if $n \geq k_2$ then $|y_n - d| < \frac{c-d}{2}$, meaning that $d - \frac{c-d}{2} < y_n < d + \frac{c-d}{2}$. Thus, if $n \geq k = \max\{k_1, k_2, j\}$ then $y_n < \frac{d+c}{2} < x_n$, which is impossible since we are given that $x_n \leq y_n$ since $n \geq j$.

□

Theorem 3.8. *Let $\{a_n\} \to a \neq 0$ and let $a_n \neq 0$ for each $n \in \mathbb{N}$. Then $\{\frac{1}{a_n}\}$ is bounded.*

Proof. We can choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|a_n - a| < \dfrac{|a|}{2}$, and by the Triangle Inequality, $|a_n| = |a - (a - a_n)| \geq |a| - |a_n - a| > \dfrac{|a|}{2}$. Hence, if $n \geq k$ then $\dfrac{1}{|a_n|} < \dfrac{2}{|a|}$. It follows that $\dfrac{1}{|a_n|} \leq \max\{\dfrac{2}{|a|}, \dfrac{1}{|a_1|}, \dfrac{1}{|a_2|}, ..., \dfrac{1}{|a_{k-1}|}\}$ for all $n \in \mathbb{N}$, so $\{\dfrac{1}{a_n}\}$ is bounded.

□

Theorem 3.9. *Let $\{a_n\} \to a$ and $\{b_n\} \to b$. Then:*

(a) $\{a_n + b_n\} \to a + b$
(b) $\{a_n b_n\} \to ab$
(c) If $b \neq 0$ and for each $n \in \mathbb{N}$, $b_n \neq 0$, then $\dfrac{a_n}{b_n} \to \dfrac{a}{b}$.
(d) $\{ca_n + db_n\} \to ca + db$

Proof. (a) Let $\epsilon > 0$. Pick $k_1 \in \mathbb{N}$ so that if $n \geq k_1$ then $|a_n - a| < \dfrac{\epsilon}{2}$ and pick $k_2 \in \mathbb{N}$ so that if $n \geq k_2$ then $|b_n - b| < \dfrac{\epsilon}{2}$. If $n \geq \max\{k_1, k_2\}$ then $|(a_n + b_n) - (a + b) \leq |a_n - a| + |b_n - b| < \epsilon$ by the Triangle Inequality.

(b) Note that $\{a_n b_n - ab\} = \{a_n b_n - a_n b + a_n b - ab\} = \{a_n(b_n - b) + b(a_n - a)\}$. Since by Theorem 3.5 we know $\{a_n\}$ is bounded and $\{b_n - b\} \to 0$ by Theorem 3.3 it follows that $\{a_n(b_n - b)\} \to 0$ by Theorem 3.4. Similarly, $\{b(a_n - a) \to 0\}$. Hence, by (a), $\{a_n b_n - ab\} \to 0$ so $\{a_n b_n\} \to ab$ by Theorem 3.3.

(c) $\{\dfrac{a_n}{b_n} - \dfrac{a}{b}\} = \{\dfrac{a_n b - ab + ab - b_n a}{b_n b}\} = \{(\dfrac{1}{b_n})(\dfrac{1}{b})(b(a_n - a) + a(b - b_n))\}$. We know $\{\dfrac{1}{bb_n}\}$ is bounded by Theorem 3.8 and $\{b(a_n - a) + a(b - b_n)\} \to 0$, so $\{\dfrac{a_n}{b_n} - \dfrac{a}{b}\} \to 0$ by Theorem 3.4, which means that $\dfrac{a_n}{b_n} \to \dfrac{a}{b}$ by Theorem 3.3.

(d) By (b) and Theorem 3.2 we know that $\{ca_n\} \to ca$ and $\{db_n\} \to db$. By (a), it follows that $\{ca_n + db_n\} \to ca + db$.

□

Part (a) of the preceding theorem is the sum rule for sequence limits, part (b) is the product rule for sequence limits, and part (c) is the quotient rule for sequence limits. It is more convenient for us to have some way to refer to (d) so we will just use ”sum rule” to refer to (d), which is a generalization of (a).

Theorem 3.10. *A sequence can converge to at most one number.*

Proof. Let $\{x_n\} \to s$ and let $\{x_n\} \to t$. By the Comparison Theorem, since $x_n \leq x_n$ for all $n \in \mathbb{N}$, we know that $s \leq t$ and $t \leq s$ so $s = t$. □

Theorem 3.11. *Let $\{x_n\} \to c$ and let $\{x_{n_i}\}$ be a subsequence of $\{x_n\}$. Then $\{x_{n_i}\} \to c$.*

Proof. By Exercise 3.12, $n_i \geq i$ for each $i \in \mathbb{N}$. Let $\epsilon > 0$. We may choose a $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - c| < \epsilon$, so if $i \geq k$ then $n_i \geq k$, and hence $|x_{n_i} - c| < \epsilon$. Thus, $\{x_{n_i}\} \to c$. □

Another proof uses the idea of the number of sequence numbers excluded from an interval. This may be more intuitive for some readers.

Proof. By Exercise 3.6, a sequence converges to a point if and only if every open interval containing that point excludes at most finitely many sequence terms. Since $\{x_n\} \to c$, for any open interval I containing c, there are only finitely many integers n so that $x_n \notin I$. Thus, there are only finitely many integers n_i so that $x_{n_i} \notin I$, so $\{x_{n_i}\} \to c$. □

Definition 19

Let $h : D \to \mathbb{R}$ be a function. We say that h is *increasing* if $h(a) < h(b)$ whenever $a < b$ and $a, b \in D$. We say h is *decreasing* if $h(a) > h(b)$ whenever $a < b$ and $a, b \in D$. We say that h is *non-decreasing* if $h(a) \leq h(b)$ whenever $a < b$ and $a, b \in D$. We say h is *non-increasing* if $h(a) \geq h(b)$ whenever $a < b$ and $a, b \in D$. A function which is either non-increasing or non-decreasing is referred to as *monotone*.

Note that since sequences are functions, they can be increasing, decreasing, non-increasing or non-decreasing or monotone as defined above. Thus, a sequence $\{x_n\}$ is non-decreasing if $x_i \leq x_j$ whenever $i < j$ and a sequence $\{x_n\}$ is non-increasing if $x_i \geq x_j$ whenever $i < j$. Likewise, a sequence $\{x_n\}$ is increasing if $x_i < x_j$ whenever $i < j$, and a sequence $\{x_n\}$ is decreasing if $x_i > x_j$ whenever $i < j$. A sequence is monotone if it is non-increasing or non-decreasing.

Theorem 3.12. *Monotone Convergence Theorem. Let $\{x_n\}$ be a bounded monotone sequence. If $\{x_n\}$ is non-decreasing then $\{x_n\}$ converges to its least upper bound. If $\{x_n\}$ is non-increasing then $\{x_n\}$ converges to its greatest lower bound.*

Proof. First, assume $\{x_n\}$ is non-decreasing, let $u = \sup(\{x_n\})$ and let $\epsilon > 0$. By the Approximation Property, there is a $k \in \mathbb{N}$ so that $u - \epsilon < x_k \leq u$. Since $\{x_n\}$ is non-decreasing, it follows that $u - \epsilon < x_k \leq x_n \leq u$ and hence $|x_n - u| < \epsilon$ for all $n \geq k$.

Next, assume $\{x_n\}$ is non-increasing, let $b = \inf(\{x_n\})$ and let $\epsilon > 0$. By the Approximation Property, there is a $k \in \mathbb{N}$ so that $b \leq x_k < b + \epsilon$. Since $\{x_n\}$ is non-increasing, it follows that $b \leq x_n \leq x_k < b + \epsilon$ and hence $|x_n - b| < \epsilon$ for all $n \geq k$. □

Definition 20

We say that a point p is a *limit point* of a set S if for every $\epsilon > 0$ there is a point $s \in S$ distinct from p so that $|p - s| < \epsilon$. In other words, p is a limit point of S if, for every $\epsilon > 0$, $(p - \epsilon, p + \epsilon) \cap S \setminus \{p\} \neq \emptyset$. If $p \in S$ and p is not a limit point of S then we refer to p as an *isolated point* of S.

The usual definition of limit point is that a point p is a limit point of a set S if every open set containing p (or, equivalently, every neighborhood of p) contains a point of S distinct from p, but it is more convenient for us to use the definition above (partly since we have not yet defined what it means for a set to be open). The fact that these two definitions are equivalent is Exercise 3.4.

Note that a point $p \in S$ is an isolated point of S if and only if p is contained in an open interval $(p - \epsilon, p + \epsilon)$ for some $\epsilon > 0$, so that $(p - \epsilon, p + \epsilon)$ contains no point of S other than p.

Theorem 3.13. *A point p is a limit point of a set A if and only if every open interval containing p contains infinitely many points of A.*

Proof. First, assume that every open interval containing p contains infinitely many points of A. Let $\epsilon > 0$. Then $(p - \epsilon, p + \epsilon)$ is an open interval containing p, and hence contains infinitely many points of A, including points distinct from p. Hence, p is a limit point of A.

Next, assume that p is a limit point of A. Suppose there is an interval (a, b) containing p and only finitely many points of A. Finite sets have first and last points, so if we let b' be the first point of A in (a, b) which is greater than p and a' be the last point of A in (a, b) which is less than p, then the interval (a', b') contains no points of A distinct from p. Hence, if we set $\epsilon = \min(|p - a'|, |p - b'|)$ then it follows that there are no points of A distinct from p which have distance less than ϵ from p, contradicting the definition of limit point.

□

Theorem 3.14. *A point p is a limit point of a set A if and only if there is a sequence $\{x_n\} \subseteq A \setminus \{p\}$ so that $\{x_n\} \to p$.*

Proof. First, assume that p is a limit point of A. Then for each $n \in \mathbb{N}$ there is some $x_n \in A$ distinct from p so that $|p - x_n| < \frac{1}{n}$. We know that $p - \frac{1}{n} < x_n < p + \frac{1}{n}$ for each $n \in \mathbb{N}$ so $\{x_n\} \to p$ by the Squeeze Theorem.

Next, assume there is a sequence $\{x_n\} \subseteq A \setminus \{p\}$ so that $\{x_n\} \to p$. Then let $\epsilon > 0$. For some $k \in \mathbb{N}$, if $n \geq k$ then $|x_n - p| < \epsilon$, so x_k is a point of A distinct from p so that $|x_k - p| < \epsilon$, and thus p is a limit point of A.

□

The preceding theorem helps us to understand one of the reasons for calling a limit point a limit point. A point p is a limit point of a set S if p is the limit of a sequence in $S \setminus \{p\}$.

Example 3.2. *(a) What are the limit points of* $(0,1]$*?*

(b) What are the limit points of $\{\frac{1}{n}\}$*?*

(c) What are the limit points of $\mathbb{Z}$*?*

(a) $[0,1]$ since every open interval about any point p in this interval intersects $(0,1]$ at points other than p.

(b) The point $\{0\}$ is the only limit point of this sequence. Since $\{\frac{1}{n}\} \to 0$ we know that 0 is a limit point of $\{\frac{1}{n}\}$ by Theorem 3.14. Since a sequence can converge to only one point, any real number p other than zero is contained in an open interval that contains at most finitely many members of $\{\frac{1}{n}\}$ by Exercise 3.6, which means that p is not a limit point of $\{\frac{1}{n}\}$ by Theorem 3.13.

(c) The set $\mathbb{Z}$ has no limit points. Let $p \in \mathbb{R}$. If there is an integer k so that $k \in (p-\frac{1}{2}, p+\frac{1}{2})$ then $p-\frac{1}{2} < k < p+\frac{1}{2}$, so $k-1 < p-\frac{1}{2} < k < p+\frac{1}{2} < k+1$. Since there are no integers between k and $k-1$ and no integers between k and $k+1$, $(k-1,k+1)\cap\mathbb{Z} = \{k\}$, which means that $(p-\frac{1}{2}, p+\frac{1}{2})$ can contain at most one integer and is therefore not a limit point of $\mathbb{Z}$ by Theorem 3.13.

Definition 21

A set U is *open* if for every $p \in U$ there is an $\epsilon > 0$ so that $(p-\epsilon, p+\epsilon) \subseteq U$. A set A is *closed* if its complement is open. If $S \subset \mathbb{R}$ then we say that U is *relatively open in* S or just *open in* S if there is an open set V so that $V \cap S = U$. Likewise, we say that a set A is *closed in* S or *relatively closed in* S if there is a closed set H in so that $H \cap S = A$. A set E that contains $(p-\epsilon, p+\epsilon)$ for some $\epsilon > 0$ is referred to as a *neighborhood* of p. A set $A \subseteq \mathbb{R}$ with open and closed sets in A as described is called a *subspace* of $\mathbb{R}$ whose topology is induced by $\mathbb{R}$ under the subspace topology. The *interior* of set S is denoted by $S^\circ = \{x \in S | (x-\epsilon, x+\epsilon) \subseteq S \text{ for some } \epsilon > 0\}$. The *closure* of a set S is denoted by $\overline{S}$ and is the set consisting of S and all limit points of S. The *boundary* of S is denoted by $\partial(S)$ and is the set of points p so that for every $\epsilon > 0$, the interval $(p-\epsilon, p+\epsilon)$ contains a point in S and a point which is not in S.

Upon first encountering terms like "open" and "closed" the reasons for the choices of words to describe open and closed sets may seem somewhat arbitrary, and it can be helpful to have some way to mentally associate these ideas with their names. One can think of a set being closed as being closed under sequence convergence, meaning that there is no way to approach a point external to the set by taking a sequence within the set converging to that point. Every point a sequence in a closed set can converge to is a point in the set.

The idea of a set being open can be thought of in terms of having freedom to vary near a point without being blocked in (so movement options are open within a certain distance without leaving the set). Each point in an open set is within an open interval contained in that set, meaning there are points in both directions (within some distance) from the point in question that remain in the set. In this manner, we don't have to be concerned with sequences from outside of the open set converging to points in the set because they can't be chosen arbitrarily closely to a point in the open set.

Thus, if sequence convergence is thought of as somehow arriving at a destination point then for closed sets, sequences inside the set can't arrive anywhere outside the set, and for open sets, sequences outside the set can't arrive anywhere inside the set.

Closed sets and open sets are complementary as sets, but they are not logically complementary. A set that is not open need not be closed, and a set that is not closed need not be open. Some sets are closed and open. A set may be open, closed, both or neither.

Theorem 3.15. *A set A is closed if and only if A contains all of its limit points.*

Proof. Let A be closed and let $p \notin A$. Then since $\mathbb{R} \setminus A$ is open, there is an $\epsilon > 0$ such that $(p-\epsilon, p+\epsilon) \subseteq \mathbb{R} \setminus A$, so p is not a limit point of A.

Let A be a set containing all of its limit points and let $p \notin A$. Since p is not a limit point of A there is an an $\epsilon > 0$ such that $(p-\epsilon, p+\epsilon) \cap A = \emptyset$, so $(p-\epsilon, p+\epsilon) \subseteq \mathbb{R} \setminus A$, and hence $\mathbb{R} \setminus A$ is open.

□

Example 3.3. *Let $A = (0,1] \cup (\mathbb{Q} \cap (3,\infty))$.*

(a) Is A open, closed, both or neither?

(b) Is $(0, \frac{1}{2}]$ open, closed or neither in A?

(c) Is $(\pi, 2\pi) \cap A$ open, closed or neither in A? Assume we know that π and 2π are irrational.

Solution. (a) A is neither open nor closed. The point 1 is not contained in an open interval which is contained in A (likewise, none of the rational numbers in $(3,\infty)$ are in the interior of A), so A is not open. The point 0 is a limit point of A and is not contained in A (the same can be said of any of the irrational numbers which are greater than three), so A is not closed.

(b) The set $(0, \frac{1}{2}] = [0, \frac{1}{2}] \cap A$, so $(0, \frac{1}{2}]$ is closed in A. Every open interval containing $\frac{1}{2}$ contains points of A which exceed $\frac{1}{2}$, which means that $(0, \frac{1}{2}]$ is not the intersection of an open set with A, so $(0, \frac{1}{2}]$ is not open in A.

(c) Since $(\pi, 2\pi)$ is open, $(\pi, 2\pi) \cap A$ is open in A. Since $\pi, 2\pi \notin A$, $[\pi, 2\pi] \cap A = (\pi, 2\pi) \cap A$ is also closed in the subspace A.

□

Example 3.4. *(a) Give an example, with proof, of a set which is both open and closed.*

(b) Give an example, with proof, of a set which is neither open nor closed.

(c) Let $S = (0,1) \cup \{3\}$. What are $\overline{S}, S^\circ$ and $\partial(S)$?

Solution. (a) $\mathbb{R}$ is both open and closed. For every $p \in \mathbb{R}$, we know $(p-1, p+1) \subset \mathbb{R}$, so $\mathbb{R}$ is open. We also know $\mathbb{R}$ is closed since it contains all points of $\mathbb{R}$ and thus all limit points of $\mathbb{R}$.

(b) $(0,1]$ is neither open nor closed. It is not open since for every $\epsilon > 0$, the interval $(1-\epsilon, 1+\epsilon)$ contains $1 + \dfrac{\epsilon}{2}$, which is not in $(0,1]$. It is not closed because 0 is a limit point of $(0,1]$ which is not contained in $(0,1]$. We can see this because for each $\epsilon > 0$, the interval $(0-\epsilon, 0+\epsilon) \cap (0,1]$ contains $\dfrac{\epsilon}{2+\epsilon}$.

(c) $\overline{S} = [0,1] \cup \{3\}$, $S^\circ = (0,1)$ and $\partial(S) = \{0,1,3\}$.

□

We did not prove (c) in the preceding example. It might be instructive for the reader to do so. It may also be worth noting that the only sets which are both open and closed as subsets of the real numbers are the set of real numbers and the empty set, though we are not going to prove that yet.

Theorem 3.16. *Let $A \subseteq \mathbb{R}$. Then A is a closed set if and only if for every $\{x_n\} \subseteq A$ so that $\{x_n\}$ converges to some point p, it is true that $p \in A$.*

Proof. Assume A is closed and $\{x_n\} \to p$. If $p \in \{x_n\}$ then $p \in A$. Otherwise, p is a limit point of A by Theorem 3.14, so $p \in A$ since A is closed.

Assume that for every $\{x_n\} \subseteq A$ so that $\{x_n\}$ converges to some point p, it is true that $p \in A$. Let p be a limit point of A. Then by Theorem 3.14, we know that there is a sequence $\{x_n\} \subseteq A \setminus \{p\}$ which converges to p. Thus, $p \in A$. □

Theorem 3.17. *If U_α is open for all $\alpha \in J$ then $\displaystyle\bigcup_{\alpha \in J} U_\alpha$ is open.*

Proof. Let $p \in \displaystyle\bigcup_{\alpha \in J} U_\alpha$. Then $p \in U_\beta$ for some $\beta \in J$, so for some $\epsilon > 0$ it follows that $(p-\epsilon, p+\epsilon) \subseteq U_\beta \subseteq \displaystyle\bigcup_{\alpha \in J} U_\alpha$, so $\displaystyle\bigcup_{\alpha \in J} U_\alpha$ is open.

□

Theorem 3.18. *If $U_1, U_2, ..., U_n$ are open sets then $\displaystyle\bigcap_{i=1}^{n} U_i$ is open.*

Proof. If $p \in \displaystyle\bigcap_{i=1}^{n} U_i$ then for each $i \leq n$ we can find $\epsilon_i > 0$ so that $(p-\epsilon_i, p+\epsilon_i) \subseteq U_i$. Hence, if we set $\epsilon = \min(\epsilon_1, \epsilon_2, ..., \epsilon_n)$ then $(p-\epsilon, p+\epsilon) \subseteq \displaystyle\bigcap_{i=1}^{n} U_i$.

□

Theorem 3.19. *If A_α is closed for all $\alpha \in J$ then $\bigcap\limits_{\alpha \in J} A_\alpha$ is closed.*

Proof. By DeMorgan's Laws, $\mathbb{R} \setminus \bigcap\limits_{\alpha \in J} A_\alpha = \bigcup\limits_{\alpha \in J} \mathbb{R} \setminus A_\alpha$, which is open by Theorem 3.17. Hence, $\bigcap\limits_{\alpha \in J} A_\alpha$ is closed.

□

Theorem 3.20. *If $A_1, A_2, ..., A_n$ are closed sets then $\bigcup\limits_{i=1}^{n} A_i$ is closed.*

Proof. By DeMorgan's Laws, $\mathbb{R} \setminus \bigcup\limits_{i=1}^{n} A_i = \bigcap\limits_{i=1}^{n} \mathbb{R} \setminus A_i$, which is open by Theorem 3.18. Hence, $\bigcup\limits_{i=1}^{n} A_i$ is closed.

□

It is not true, in general, the the union of closed sets is closed or the intersection of open sets is open if the collections of sets are infinite. For instance $\bigcap\limits_{n=1}^{\infty} (-\frac{1}{n}, \frac{1}{n}) = \{0\}$, which is not open, and $\bigcup\limits_{x \in (0,1)} \{x\} = (0, 1)$ which is not closed.

Theorem 3.21. *Let S be a set which is bounded above and let $l = \sup(S)$. Then there is a sequence of points $\{x_n\} \subseteq S$ which converges to l. If S is bounded below and $l = \inf(S)$ then there is also a sequence of points $\{x_n\} \subseteq S$ which converges to l.*

Proof. First, assume $l = \sup(S)$. By the Approximation Property, for each $n \in \mathbb{N}$ we can choose $x_n \in S$ so that $l - \frac{1}{n} < x_n \leq l$. By the Squeeze Theorem, we know that $\{x_n\} \to l$.

Next, assume $l = \inf(S)$. By the Approximation Property, for each $n \in \mathbb{N}$ we can choose $x_n \in S$ so that $l \leq x_n < l + \frac{1}{n}$. By the Squeeze Theorem, we know that $\{x_n\} \to l$. □

Theorem 3.22. *Let A be a closed set. If A is bounded below then A has a first point. If A is bounded above then A has a last point.*

Proof. If A is bounded above with $l = \sup(A)$ or A is bounded below with $l = \inf(A)$, then, in either case, by Theorem 3.21, there is a sequence $\{x_n\} \subseteq A$ which converges to l. Hence, by Theorem 3.16 we know that $l \in A$. □

The following theorem is important, and we include three proofs for readers who might find one proof strategy to be more intuitive than the others.

Theorem 3.23. *The Bolzano-Weierstrass Theorem. Every bounded sequence has a convergent subsequence.*

Proof. Let $\{x_n\}$ be a bounded sequence. We first show that $\{x_n\}$ has a monotone subsequence. Let $S = \{n \in \mathbb{N} | x_i \leq x_n$ for at most finitely many integers i$\}$. If S is infinite then let n_1 be the first element of S and note that since S is infinite and all but finitely many elements of S exceed n_1 by definition of S, we can find $n_2 > n_1$ so that $n_2 \in S$ and $x_{n_2} > x_{n_1}$. Similarly, if we have chosen $n_1 < n_2 < ... < n_k$ so that each $n_i \in S$ and $x_{n_1} < x_{n_2} < ... < x_{n_k}$ then we can find $n_{k+1} \in S$ so that $n_{k+1} > n_k$ and $x_{n_{k+1}} > x_{n_k}$ so the subsequence $\{x_{n_i}\}$ is increasing.

If S is finite then let n_1 be the first integer exceeding all elements of S. Then by definition, since n_1 is not an element of S, we know there are infinitely many integers i so that $x_i \leq x_{n_1}$ so we can pick $n_2 > n_1$ so that $x_{n_2} \leq x_{n_1}$. Similarly, if we have chosen $n_1 < n_2 < ... < n_k$ so that $x_{n_1} \geq x_{n_2} \geq ... \geq x_{n_k}$ then we can find $n_{k+1} > n_k$ so that $x_{n_{k+1}} \leq x_{n_k}$ so the subsequence $\{x_{n_i}\}$ is non-increasing.

By Theorem 3.12, every monotone bounded sequence converges. Hence, $\{x_n\}$ has a convergent subsequence $\{x_{n_i}\}$.

□

The following proof is slightly less brief, but it is easier to draw a picture of, which is often helpful.

Proof. Since $\{x_n\}$ is bounded, we can find lower and upper bounds a_1 and b_1 for this sequence so that $\{x_n\} \subset [a_1, b_1]$. Then if $x_n \in [a_1, \frac{a_1+b_1}{2}]$ for infinitely many integers n then we set $[a_2, b_2] = [a_1, \frac{a_1+b_1}{2}]$. Otherwise, since there are infinitely many natural numbers, it follows that $x_n \in [\frac{a_1+b_1}{2}, b_1]$ for infinitely many integers n, and we set $[a_2, b_2] = [\frac{a_1+b_1}{2}, b_1]$. If we have chosen nested intervals $[a_j, b_j]$ for all $j \leq k$ so that each $[a_j, b_j]$ contains infinitely many terms of $\{x_n\}$ and $b_j - a_j = \frac{b_1 - a_1}{2^{j-1}}$, and $a_1 \leq a_2 \leq ...a_k \leq b_k \leq b_{k-1} \leq ... \leq b_1$ then if $x_n \in [a_k, \frac{a_k+b_k}{2}]$ for infinitely many integers n then we set $[a_{k+1}, b_{k+1}] = [a_k, \frac{a_k+b_k}{2}]$. Otherwise, set $[a_{k+1}, b_{k+1}] = [\frac{a_k+b_k}{2}, b_k]$, and note that the aforementioned properties hold for $j \leq k+1$.

Choose $x_{n_1} \in [a_1, b_1]$. If we have chosen $x_{n_i} \in [a_i, b_i]$ for $i \leq k$ so that $n_1 < n_2... < n_k$, then since $x_n \in [a_{k+1}, b_{k+1}]$ for infinitely many integers n, we can choose $x_{n_{k+1}} \in [a_{k+1}, b_{k+1}]$ so that $n_{k+1} > n_k$. Then $\{x_{n_i}\}$ is a subsequence of $\{x_n\}$. Since $\{a_n\}$ is increasing and bounded above by every b_i, it follows that $\{a_n\}$ converges to $p = \sup(\{a_n\})$. We know that $\{b_n - a_n\} \to 0$ and hence $\{b_n\} \to p$ by exercise 3.8. Hence, by the Squeeze Theorem, $\{x_{n_i}\} \to p$.

□

We give another proof which is somewhat more topological (more based on open and closed sets and limit points).

Proof. Suppose $\{x_n\}$ has no limit point. Then every subset of $\{x_n\}$ has no limit point by Exercise 3.14 and is therefore closed. By Theorem 3.22, we can find n_1 so that x_{n_1} is the least element of $\{x_n\}$. Likewise, $\{x_i | i > n_1\}$ has a least element $x_{n_2} \geq x_{n_1}$ where $n_2 > n_1$. Then choose x_{n_3} to be the least element of $\{x_i | i > n_2\}$. Continuing in this manner we construct a non-decreasing bounded subsequence $\{x_{n_i}\}$ of $\{x_n\}$ which converges by Theorem 3.12.

If $\{x_n\}$ has a limit point p then choose $n_1 \in \mathbb{N}$ so that $x_{n_1} \in (p-1, p+1)$. Assume we have chosen $n_1 < n_2 < ... < n_k$ so that each $x_{n_i} \in (p - \frac{1}{i}, p + \frac{1}{i})$ for all positive integers $i \leq k$. Since p is a limit point of $\{x_n\}$ we know $(p-\frac{1}{k+1}, p+\frac{1}{k+1})$ contains x_i for infinitely many integers i so we can choose $n_{k+1} > n_k$ so that $x_{n_{k+1}} \in (p - \frac{1}{k+1}, p + \frac{1}{k+1})$. The subsequence $\{x_{n_i}\}$ thus chosen converges to p by the Squeeze Theorem.

□

It may be worth noting that in the case where $\{x_n\}$ has no limit points in the last proof, the sequence $\{x_n\}$ is constant after some point (why would this be the case?)

Definition 22

We say that $\{x_n\}$ is a *Cauchy sequence* if for every $\epsilon > 0$ there is an $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|x_n - x_m| < \epsilon$.

Theorem 3.24. *Let $\{x_n\}$ be a Cauchy sequence. Then $\{x_n\}$ is bounded.*

Proof. Choose $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|x_n - x_m| < 1$. Then, if we let $m = \min\{x_1, x_2, ..., x_{k-1}, x_k - 1\}$ and let $M = \max\{x_1, x_2, ..., x_{k-1}, x_k + 1\}$, by definition of minimum and maximum we see that if $i \leq k$ then $m \leq x_i \leq M$. Likewise, if $i \geq k$ then since $|x_i - x_k| < 1$ it follows that $m \leq x_i \leq M$, so $\{x_n\}$ is bounded.

□

Theorem 3.25. *Let $\{x_n\}$ be a sequence of real numbers. Then $\{x_n\}$ converges if and only if $\{x_n\}$ is a Cauchy sequence.*

Proof. First, assume $\{x_n\} \to p$ and let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - p| < \frac{\epsilon}{2}$. Then if $n, m \geq k$ it follows that $|x_n - x_m| = |x_n - p + p - x_m| \leq |x_n - p| + |x_m - p| < \epsilon$. Hence, $\{x_n\}$ is a Cauchy sequence.

Next, assume $\{x_n\}$ is a Cauchy sequence. Then by Theorem 3.24 we know $\{x_n\}$ is bounded, so by the Bolzano Weierstrass Theorem this sequence has a convergent subsequence $\{x_{n_i}\} \to c$. Choose k_1 so that if $i \geq k_1$ then $|x_{n_i} - c| < \frac{\epsilon}{2}$ and k_2 so that if $i, j \geq k_2$ then $|x_i - x_j| < \frac{\epsilon}{2}$. By exercise 3.12 it then follows that if $i \geq \max\{k_1, k_2\}$ then $|x_i - c| \leq |x_i - x_{n_k}| + |x_{n_k} - c| < \epsilon$. Thus, $\{x_n\} \to c$. □

The main value of a Cauchy sequence is to look at sequences that in some sense should converge to something. In subspaces of the real numbers which are not closed it is possible to find Cauchy sequences that do not converge. The points to which they should converge are missing from the space. A metric space where every Cauchy sequence converges is called a *complete* metric space, but we will not be discussing metric spaces in this text. Essentially, the completeness axiom of the real numbers causes every Cauchy sequence to converge. In fact, an equivalent axiom to the completeness axiom would be to state that in the real numbers every Cauchy sequence converges.

Definition 23

Let $\{a_n\}$ be a sequence. For any natural number k we say that the subsequence $\{a_{n+k}\}$ is a *tail* of the sequence $\{a_n\}$.

The following theorems can be helpful when using series convergence tests. They are not needed right away, so their proofs are left as exercises.

Theorem 3.26. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\} \to L$ if and only if the sequence tail $\{a_{n+k}\} \to L$.*

Theorem 3.27. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\}$ is bounded above if and only if the sequence tail $\{a_{n+k}\}$ is bounded above, and $\{a_n\}$ is bounded below if and only if the sequence tail $\{a_{n+k}\}$ is bounded below.*

Many ideas in calculus involve sequences that diverge in a specific way, which can be referred to as diverging to infinity or negative infinity. Alternately, we can say such sequences converge to infinity or negative infinity if we think of infinity and negative infinity as extended real numbers. More is discussed about extended real numbers and infinite limits for functions in the Supplementary Materials. We will just address a few theorems that come up in the study of series and integration here. While the term "converge to infinity" is used, a series that converges to infinity is divergent (it does not converge) so this terminology can be confusing.

Definition 24

We say that $\{x_n\} \to \infty$ (respectively $-\infty$), also written $\lim\limits_{n\to\infty} x_n = \infty$ (respectively $-\infty$), if, for every $M \in \mathbb{R}$ there is an integer $k \in \mathbb{N}$ so that if $n \geq k$ then $x_n > M$ (respectively $x_n < M$).

Theorem 3.28. *If $\{x_n\} \to \infty$ and $\{y_n\}$ is bounded below then $\{x_n + y_n\} \to \infty$. If $\{x_n\} \to -\infty$ and $\{y_n\}$ is bounded above then $\{x_n + y_n\} \to -\infty$.*

Proof. Let $M > 0$. If $\{x_n\} \to \infty$ and $\{y_n\}$ is bounded below by B then we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $x_n > M - B$, so $x_n + y_n > M$, which means $\{x_n + y_n\} \to \infty$.

Let $M < 0$. If $\{x_n\} \to -\infty$ and $\{y_n\}$ is bounded above by B then we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $x_n < M - B$, so $x_n + y_n < M$, which means $\{x_n + y_n\} \to -\infty$. □

Theorem 3.29. *Let $\{x_n\}$ be bounded and let $\{y_n\} \to \pm\infty$. Then $\{\frac{x_n}{y_n}\} \to 0$.*

Proof. Let $\epsilon > 0$. Choose M so that $|x_n| < M$ for all $n \in \mathbb{N}$. Choose $k \in \mathbb{N}$ so that if $\{y_n\} \to \infty$ then $y_n > \frac{M}{\epsilon}$ and if $\{y_n\} \to -\infty$ then $y_n < \frac{-M}{\epsilon}$. If $n \geq k$ then $|\frac{x_n}{y_n}| \leq |\frac{\pm\epsilon}{M}||M| = \epsilon$, so $\{\frac{x_n}{y_n}\} \to 0$. □

Apart from showing that a finite set has a first and last point, it is not essential to cover the remaining material in this section to achieve a coherent understanding of advanced calculus (this material is not used much after this point), though it would be good to observe that the real numbers in an interval consisting of more than one point are uncountable, meaning that they cannot be listed as the range of a sequence.

A section in the Supplementary Materials chapter proves what are probably the main theorems of interest about cardinality as the subject pertains to advanced calculus. Below, we just mention a few theorems that are more fundamental (and are thus included in the main body).

Optional Content: Cardinality

We often wish to talk about the number of elements in a set, and we would like to use ideas like the Pigeon Hole Principle in an argument or infinite variations of this idea, but if we wish to have access to these tools then we will need to formalize more. A more detailed and interesting development of this topic is found in the Supplementary Materials section.

For those who wish to get a sense for the main results of interest in this discussion without moving further into the topic, we include arguments for why finite sets have first and last points, why the set of real numbers within any interval containing more than one point is uncountable, and also why the rational numbers are countable.

Definition 25

We say $|A| = |B|$ or that sets A, B have the same *cardinality* if there is a one to one and onto mapping from A to B. Let $\{1, 2, 3, ..., n\}$ denote $\{i \in \mathbb{N} | i \leq n\}$. If $|A| = |\{1, 2, .., n\}|$ for some natural number n, then we say $|A| = n$. If $|A| = n$ for some non-negative integer n then we say that A is *finite*. If A is non-empty and not finite then we say that A is *infinite*. If $|A| = |\mathbb{N}|$ then we say that A is *countably infinite*. If A is countably infinite or finite then we say that A is *countable*. If A is not countable then we say that A is *uncountable*.

Theorem 3.30. *Let S be a finite non-empty subset of $\mathbb{R}$. Then S has a first and last point.*

Proof. We proceed by induction on the cardinality n of S. If S has one point then this is its first and last point. Assume that every set of cardinality k has a first and last point for some $k \in \mathbb{N}$. Then let $|S| = k+1$ and let $f : \{1, 2, ..., k+1\} \to S$ be one to one and onto. Let $g : \{1, 2, 3, ..., k\} \to S \setminus \{f(k+1)\}$ be defined by $g(i) = f(i)$. Then g is one to one and onto since f is one to one and onto (because different points map to different images under f and hence under g, and every point in $S \setminus \{f(k+1)\}$ is mapped to by a point of $\{1, 2, 3, ..., k\}$ under f and therefore under g), which means $|S \setminus \{f(k+1)\}| = k$. Then the range of g has a first point m and a last point M by the induction hypothesis. Thus, for all $1 \leq i \leq k+1$ it is true that $\min\{m, f(k+1)\} \leq f(i) \leq \max\{M, f(k+1)\}$, so S has a first point and a last point. By induction, it follows that every finite set has a first point and a last point.

□

While there are interesting proofs that the real numbers are uncountable using decimals which are more intuitive than the one below (such a proof is developed in the Supplementary Section), we can still prove that the real numbers are uncountable using only the theorems we have developed thus far. The disadvantage to using the decimal argument is that to make the argument rigorous we must prove theorems about decimal representations of real numbers (which are also in the Supplementary Materials).

Theorem 3.31. *Let S be a subset of $\mathbb{R}$ containing the interval (a, b). Then S is uncountable. In particular, $\mathbb{R}$ is uncountable.*

Proof. Suppose S is countable. We know S is non-empty since $a < b$. If S is finite there is a one to one and onto function $g : \{1, 2, 3, ...n\} \to S$ for some $n \in \mathbb{N}$, in which case we can extend g to a function $f : \mathbb{N} \to S$ which is onto by setting $f(i) = g(1)$ if $i > n$ and $f(i) = g(i)$ if $i \leq n$. Thus, whether S is finite or countably infinite we can find a function $f : \mathbb{N} \to S$ which is onto. Setting $f(i) = x_i$ for each $i \in \mathbb{N}$ we can write $S = \{x_1, x_2, x_3, ...\}$.

Choose n_1 so that $x_{n_1} \in (a, b)$. Let n_2 be the first positive integer so that $x_{n_1} < x_{n_2} < b$. Let n_3 be the first positive integer so that $x_{n_3} \in (x_{n_1}, x_{n_2})$. In general, if n_i has been chosen for $1 \leq i \leq k$ then let n_{k+1} be the first positive integer so that $x_{n_{k+1}}$ is between $x_{n_{k-1}}$ and x_{n_k}. Note that $x_{n_1} < x_{n_3} < x_{n_5} < ...$ and $x_{n_2} > x_{n_4} > ...$ and that if j is even and k is odd then $x_{n_j} > x_{n_k}$. Then by Theorem 3.12, we know that $\{x_{n_{2i+1}}\} \to u = \sup\{x_{n_{2i+1}}\}$, where $u < x_{n_j}$ if j is even and $u > x_{n_j}$ if j is odd by the Comparison Theorem, which means $u \notin \{x_{n_i}\}$. Since $u \in (a, b)$ and f is onto, there is an integer s so that $u = x_s$. But then if $u \neq x_{n_i}$ for any $i \leq s$, by construction we would have chosen $x_{n_s} = u$, which is a contradiction to the fact that $u \notin \{x_{n_i}\}$. Hence, S is uncountable.

□

We next wish to establish that $\mathbb{Q}$ is countable. Our first argument for the countability of $\mathbb{Q}$ only works if the reader is willing to accept the existence of the function based on the diagram below, and is not rigorous. We are not referring to this as a proof because of the missing details, but it is still instructive.

Explanation for why the set $\mathbb{Q}$ of rational numbers is countably infinite:

By setting $f(n)$ to be the nth rational number encountered by the following arrow path which has not yet been mapped to by an earlier natural number, we get a one to one and onto function from $\mathbb{N}$ to $\mathbb{Q}$, meaning that $\mathbb{Q}$ is countably infinite.

$$\begin{tikzcd}
\dots & -3 & -2 \arrow[d] & -1 \arrow[l] & 0 \arrow[r] & 1 \arrow[d] & 2 \arrow[r] & 3 \arrow[d] & \dots \\
\dots & \frac{-3}{2} & \frac{-2}{2} \arrow[d] & \frac{-1}{2} \arrow[u] & \frac{0}{2} \arrow[l] & \frac{1}{2} \arrow[l] & \frac{2}{2} \arrow[u] & \frac{3}{2} & \dots \\
\dots & \frac{-3}{3} & \frac{-2}{3} \arrow[r] & \frac{-1}{3} \arrow[r] & \frac{0}{3} \arrow[r] & \frac{1}{3} \arrow[r] & \frac{2}{3} \arrow[u] & \frac{3}{3} & \dots
\end{tikzcd}$$

To make the preceding argument rigorous is certainly possible and could be achieved by describing in words exactly how images that come from the arrow diagram are chosen (with a properly defined algorithm or formula instead of a picture) and then proving that the algorithm for such choices creates a one to one and onto function. Since this description seems to be a bit awkward, we use the procedure in the following argument to formalize this theorem instead. Another (nicer) development of this result is included in the Supplementary Materials section, but the argument below requires no further development.

Theorem 3.32. *The set $\mathbb{Q}$ of rational numbers is countable.*

Proof. We define $f : \mathbb{N} \to \mathbb{Q}$ inductively as follows. Let $f(1) = 0$, $f(1) = 1$ and $f(2) = -1$. If $f(i)$ has been defined for all $\{i \in \mathbb{N} | 1 \leq i \leq k\}$ for some $k \geq 2$ then we let s be the least natural number so that there is a rational number $\frac{p}{q} \neq 0$ having the property that $|p| + |q| = s$ and $f(i) \neq \frac{p}{q}$ for any $i \leq k$. Pick any $p, q \in \mathbb{Z}$ so that $f(i) \neq \frac{p}{q}$ for any $i \leq k$ and $|p| + |q| = s$ and assign $f(k+1) = \frac{p}{q}$.

Since each choice of $f(k+1)$ is a rational number which is not $f(i)$ for any $i \leq k$, it follows that f is one to one. Note that for any integer $s \geq 2$, there are only $2s - 2$ distinct pairs of non-zero numbers $a \in \mathbb{N}, b \in \mathbb{Z}$ having the property that $|a| + |b| = s$ (specifically, $(\pm 1, s-1)$, $(\pm 2, s-2)$,... $(\pm(s-1), 1)$). Hence, all rational numbers $\frac{a}{b}$ having the property that $|a|+|b| = s$ have been mapped to by positive integers i so that $i \leq s(s-1)$ by Theorem 2.2 because all such rational numbers have been chosen as images of integers within the first $1 + 2 + 4 + 6 + ... + 2s - 2$ natural numbers. Since all rational numbers are in the range of f, it follows that $\mathbb{Q}$ is countably infinite.

□

Exercises:

Exercise 3.1. *Let $\{x_n\}$ be a sequence. Then $\{x_n\} \to c$ if and only if $\{|x_n - c|\} \to 0$.*

Exercise 3.2. *An open interval (a, b) is an open set.*

Exercise 3.3. *A closed interval $[a, b]$ is closed set.*

Exercise 3.4. *Let $S \subseteq \mathbb{R}$ and let $p \in \mathbb{R}$. Then p is a limit point of S if and only if every open set containing p contains a point of S distinct from p.*

Note that the condition in the preceding exercise is usually used as the definition of a point p being a limit point of a set S in general.

Exercise 3.5. *Let $l = \sup(S)$. If $l \notin S$ then there is an increasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to l$.*

Let $b = \inf(T)$. If $b \notin T$ then there is a decreasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to b$.

Exercise 3.6. *Let $\{x_n\}$ be a sequence of real numbers. Then $\{x_n\} \to p$ if and only if every open interval containing p excludes x_n for at most finitely many positive integers n. This is true if and only if for every $\epsilon > 0$ there are at most finitely many positive integers n so that $x_n \notin (p - \epsilon, p + \epsilon)$.*

Exercise 3.7. *Any interval whose right endpoint is b has a supremum equal to b. Any interval whose left endpoint is a has a infimum equal to a.*

Exercise 3.8. *If $\{a_n\} \to a$ and $\{a_n + b_n\} \to a + b$, then $\{b_n\} \to b$.*

Exercise 3.9. *If $\{a_n\} \to a \neq 0$ and $\{a_n b_n\} \to ab$, then $\{b_n\} \to b$.*

Exercise 3.10. *Find an example of sequences $\{x_n\}$ and $\{y_n\}$ which both diverge, such that $\{x_n y_n\}$ converges.*

Exercise 3.11. *Let r be a real number. There is a sequence of rational numbers converging to r.*

Exercise 3.12. *Let $\{x_{n_i}\}$ be a subsequence of $\{x_n\}$. Then $n_i \geq i$ for each $i \in \mathbb{N}$.*

Exercise 3.13. *If $|r| < 1$ then $\{r^n\} \to 0$.*

Exercise 3.14. *Let p be a limit point of a set A and let $A \subseteq B$. Then p is a limit point of B.*

Exercise 3.15. *Let p be a limit point of $A \cup B$. Then either p is a limit point of A or p is a limit point of B.*

Exercise 3.16. *Let E be a set and let $A \subseteq E$. Then $E \setminus A$ is open in E if and only if A is closed in E.*

Exercise 3.17. *Let A_i be closed, non-empty and bounded for each natural number i, so that $A_1 \supseteq A_2 \supseteq A_3 \ldots.$ Then $\displaystyle\bigcap_{i=1}^{\infty} A_i$ is non-empty.*

Exercise 3.18. *Let S be a bounded infinite set. Then S has a limit point.*

Exercise 3.19. *The sets $\mathbb{Z}, \mathbb{R}, \mathbb{Q}$ are infinite sets.*

Exercise 3.20. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\} \to L$ if and only if the sequence tail $\{a_{n+k}\} \to L$.*

Exercise 3.21. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\}$ is bounded above if and only if the sequence tail $\{a_{n+k}\}$ is bounded above, and $\{a_n\}$ is bounded below if and only if the sequence tail $\{a_{n+k}\}$ is bounded below.*

Hints:

Hint to Exercise 3.1. *Let $\{x_n\}$ be a sequence. Then $\{x_n\} \to c$ if and only if $\{|x_n - c|\} \to 0$.*

Use Theorem 3.3.

Hint to Exercise 3.2. *An open interval (a, b) is an open set.*

Take an arbitrary point p in (a, b) and find an ϵ small enough so that $(p-\epsilon, p+\epsilon) \subseteq (a, b)$.

Hint to Exercise 3.3. *Let $S \subseteq \mathbb{R}$ and let $p \in \mathbb{R}$. Then p is a limit point of S if and only if every open set containing p contains a point of S distinct from p.*

Use the result of Exercise 3.2.

Hint to Exercise 3.4. *A closed interval $[a, b]$ is closed set.*

Show that open rays are open.

Hint to Exercise 3.5. *Let $l = \sup(S)$. If $l \notin S$ then there is an increasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to l$.*

Let $b = \inf(T)$. If $b \notin T$ then there is a decreasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to b$.

Parallel the proof of Theorem 3.21

Hint to Exercise 3.6. *Let $\{x_n\}$ be a sequence of real numbers. Then $\{x_n\} \to p$ if and only if every open interval containing p excludes x_n for at most finitely many positive integers n. This is true if and only if for every $\epsilon > 0$ there are at most finitely many positive integers n so that $x_n \notin (p - \epsilon, p + \epsilon)$.*

Write the definition of convergence, and note that the first k sequence terms is a finite set. For the other direction, remember that if there is a finite number of sequence terms excluded by $(p - \epsilon, p + \epsilon)$ then the corresponding indices are a finite set, meaning there is a last index of an excluded point.

Hint to Exercise 3.7. *Any interval whose right endpoint is b has a supremum equal to b. Any interval whose left endpoint is a has a infimum equal to a.*

Use the definition of interval, supremum and the fact that there is a real number between any two numbers (specifically, a rational number has been shown to be between any two numbers).

Hint to Exercise 3.8. *If $\{a_n\} \to a$ and $\{a_n + b_n\} \to a + b$, then $\{b_n\} \to b$.*

Use the sum rule for sequence limits. Remember that you cannot assume that $\{b_n\}$ converges.

Hint to Exercise 3.9. *If $\{a_n\} \to a \neq 0$ and $\{a_n b_n\} \to ab$, then $\{b_n\} \to b$.*

Use the product rule for sequence limits. Remember that you cannot assume that $\{b_n\}$ converges.

Hint to Exercise 3.10. *Find an example of sequences $\{x_n\}$ and $\{y_n\}$ which both diverge, such that $\{x_n y_n\}$ converges.*

There are examples where $\{x_n\}$ diverges and $\{(x_n)^2\}$ converges. It might be easier to think of such a sequence. You want one sequence to move points in the other close to the remaining points of that sequence when the sequences are multiplied together.

Hint to Exercise 3.11. *Let r be a real number. There is a sequence of rational numbers converging to r.*

Start with a sequence converging to r and use the fact that there is a rational number between any two points and apply the Squeeze Theorem.

Hint to Exercise 3.12. *Let $\{x_{n_i}\}$ be a subsequence of $\{x_n\}$. Then $n_i \geq i$ for each $i \in \mathbb{N}$.*

Use induction.

Hint to Exercise 3.13. *If $|r| < 1$ then $\{r^n\} \to 0$.*

First, explain why $\{|r|^n\}$ converges (what kind of sequence is it?). Then consider the subsequence $\{|r|^{n+1}\}$ of $\{|r|^n\}$. What does this subsequence converge to according to the product rule for limits? What does it converge to according to the theorem that a subsequence converges to the same point as the sequence it is a subsequence of?

Hint to Exercise 3.14. *Let p be a limit point of a set A and let $A \subseteq B$. Then p is a limit point of B.*

Use the definition of limit point and subset.

Hint to Exercise 3.15. *Let p be a limit point of $A \cup B$. Then either p is a limit point of A or p is a limit point of B.*

Assume p is not a limit point of A, and explain why it must be a limit point of B.

Hint to Exercise 3.16. *Let E be a set and let $A \subseteq E$. Then $E \setminus A$ is open in E if and only if A is closed in E.*

Use the definitions of open and closed in E. If there is an open set U so that $U \cap E = A$ then what does this say about $\mathbb{R} \setminus U \cap E$?

Hint to Exercise 3.17. *Let A_i be closed, non-empty and bounded for each natural number i, so that $A_1 \supseteq A_2 \supseteq A_3 \dots$. Then $\bigcap_{i=1}^{\infty} A_i$ is non-empty.*

Take a point in each A_i. The points chosen form a bounded sequence. What is known about bounded sequences? If a set is closed, what is known about the limit of a sequence of points from that set?

Hint to Exercise 3.18. *Let S be a bounded infinite set. Then S has a limit point.*

Use the Bolzano-Weierstrass Theorem.

Hint to Exercise 3.19. *The sets $\mathbb{Z}, \mathbb{R}, \mathbb{Q}$ are infinite sets.*

Do any of these sets have last points?

Hint to Exercise 3.20. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\} \to L$ if and only if the sequence tail $\{a_{n+k}\} \to L$.*

Try looking at what happens in the sequence k indices later and compare that to corresponding sequence points in the tail of the sequence.

Hint to Exercise 3.21. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\}$ is bounded above if and only if the sequence tail $\{a_{n+k}\}$ is bounded above, and $\{a_n\}$ is bounded below if and only if the sequence tail $\{a_{n+k}\}$ is bounded below.*

Consider that the difference between a sequence and its tail is only finitely many points, and finite sets have first and last points.

Solutions:

Solution to Exercise 3.1. *Let $\{x_n\}$ be a sequence. Then $\{x_n\} \to c$ if and only if $\{|x_n - c|\} \to 0$.*

Proof. We know $\{x_n\} \to c$ if and only if $\{x_n - c\} \to 0$ by Theorem 3.3, which is true if and only if, for every $\epsilon > 0$ there is an integer k so that if $n \geq k$ then $|x_n - c| < \epsilon$. Since $|x_n - c| = ||x_n - c| - 0|$, this is true if and only if $\{|x_n - c|\} \to 0$. □

Solution to Exercise 3.2. *An open interval (a, b) is an open set.*

Proof. Let $p \in (a, b)$ and let $\epsilon = \min\{p - a, b - p\}$. Then $p - \epsilon \geq p - (p - a) = a$ and $p + \epsilon \leq p + (b - p) = b$. Thus, $(p - \epsilon, p + \epsilon) \subseteq (a, b)$. Hence, (a, b) is open. □

Solution to Exercise 3.3. *A closed interval $[a, b]$ is closed set.*

Proof. By Exercise 3.2, we know that $(b, b + n)$ and $(a - n, a)$ are open for each natural number n. Hence, $U = \bigcup_{n=1}^{\infty} (a - n, a) \cup (b, b + n)$ is open by Theorem 3.17. If $M > b$ then since $\mathbb{N}$ is unbounded, there is some $n \in \mathbb{N}$ so that $n > M - b$, so $b + n > M$, which means $M \in U$. By a similar argument, if $M < a$ then $M \in U$. Hence $U = \mathbb{R} \setminus [a, b]$, which means that $[a, b]$ is closed. □

Solution to Exercise 3.4. *Let $S \subseteq \mathbb{R}$ and let $p \in \mathbb{R}$. Then p is a limit point of S if and only if every open set containing p contains a point of S distinct from p.*

Proof. Assume p is a limit point of S. Let U be an open set containing p. Then for some $\epsilon > 0$ we know $(p - \epsilon, p + \epsilon) \subseteq U$, which means that there is some $q \neq p$ so that $q \in (p - \epsilon, p + \epsilon) \cap S \subseteq (U \cap S)$ (since p is a limit point of S).

Assume that every open set containing p contains a point of S distinct from p. Let $\epsilon > 0$. Since the interval $(p - \epsilon, p + \epsilon)$ is an open set by Exercise 3.2, it follows that $(p - \epsilon, p + \epsilon)$ contains a point of S distinct from p, which implies that p is a limit point of S. □

Solution to Exercise 3.5. *Let $l = \sup(S)$. If $l \notin S$ then there is an increasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to l$.*

Let $b = \inf(T)$. If $b \notin T$ then there is a decreasing sequence $\{x_n\} \subseteq S$ so that $\{x_n\} \to b$.

Proof. Let $l = \sup(S)$, where $l \notin S$. By the Approximation Property we can choose $x_1 \in S \cap (l-1, l]$. Since $l \notin S$, $l-1 < x_1 < l$. Likewise, we can choose $x_2 \in (\max(x_1, l-1), l)$. Continuing, we choose each x_n so that $\max(l - \frac{1}{n}, x_{n-1}) < x_n < l$ for all $n > 1$. Then $\{x_n\}$ is increasing and converges to l by the Squeeze Theorem.

Let $b = \inf(S)$, where $b \notin S$. By the Approximation Property we can choose $x_1 \in S \cap [b, b+1)$. Since $b \notin S$, $b+1 > x_1 > b$. Likewise, we can choose $x_2 \in (b, \min\{x_1, b+1\})$. Continuing, we choose each x_n so that $\min\{b + \frac{1}{n}, x_{n-1}\} > x_n > b$ for all $n > 1$. Then $\{x_n\}$ is decreasing and converges to l by the Squeeze Theorem.

□

Solution to Exercise 3.6. *Let $\{x_n\}$ be a sequence of real numbers. Then $\{x_n\} \to p$ if and only if every open interval containing p excludes x_n for at most finitely many positive integers n. This is true if and only if for every $\epsilon > 0$ there are at most finitely many positive integers n so that $x_n \notin (p - \epsilon, p + \epsilon)$.*

Proof. First, note that if $p \in (a, b)$ then by setting $\epsilon = \min\{|p - a|, |p - b|\}$, it follows that $(p - \epsilon, p + \epsilon) \subseteq (a, b)$. Thus, if $x_n \notin (a, b)$ for infinitely many integers n then $x_n \notin (p - \epsilon, p + \epsilon)$ for infinitely many integers n. Likewise, if there is some $\epsilon > 0$ so that $x_n \notin (p - \epsilon, p + \epsilon)$ for infinitely many integers n, then there is an open interval (namely $(p - \epsilon, p + \epsilon)$) the excludes x_n for infinitely many integers n.

Assume $\{x_n\} \to p$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - p| < \epsilon$. Then if $n \geq k$ we know $x_n \in (p - \epsilon, p + \epsilon)$. Thus, the set S of integers i so that $x_i \notin (p - \epsilon, p + \epsilon)$ is a subset of $\{1, 2, ..., k - 1\}$, which is finite, and thus S is finite.

Assume that, for every $\epsilon > 0$ there are only finitely many integers n so that $x_n \notin (p - \epsilon, p + \epsilon)$. Then there is a last integer $k - 1$ so that $x_{k-1} \notin (p - \epsilon, p + \epsilon)$. Hence, if $n \geq k$ then $x_n \in (p - \epsilon, p + \epsilon)$, so $|x_n - p| < \epsilon$, which means that $\{x_n\} \to p$. □

Solution to Exercise 3.7. *Any non-empty interval whose right endpoint is b has a supremum equal to b. Any interval whose left endpoint is a has a infimum equal to a.*

Proof. If the right end point of an interval I is b then for some number $a < b$, $a \in I$ so all numbers between a and b are in I. Thus, if $c < b$ then there is a rational number q between $\max(a, c)$, and b, so $q \in I$ and c is not an upper bound for I. Since I contains no points greater than its right end point, b is the least upper bound for I.

If b is the left end point of an interval I then as before, we can find a number in I less than any number exceeding b, and it follows that b is the greatest lower bound of I.

□

Solution to Exercise 3.8. *If $\{a_n\} \to a$ and $\{a_n + b_n\} \to a + b$, then $\{b_n\} \to b$.*

Proof. By the sum rule (and product rule) we know that $\{(a_n + b_n) - a_n\} \to a + b - a$, which means that $\{b_n\} \to b$

□

Solution to Exercise 3.9. *If $\{a_n\} \to a \neq 0$ and $\{a_n b_n\} \to ab$, then $\{b_n\} \to b$.*

Proof. By product and quotient rules we know that $\{\frac{1}{a_n}\} \to \frac{1}{a}$, which means that $\{b_n\} = \{\frac{1}{a_n}(a_nb_n)\} \to (ab)(\frac{1}{a}) = b$. □

Solution to Exercise 3.10. *Find an example of sequences $\{x_n\}$ and $\{y_n\}$ which both diverge, such that $\{x_ny_n\}$ converges.*

Proof. We will use $x_n = y_n = (-1)^n$. Then $x_ny_n = 1$ for each $n \in \mathbb{N}$ so $\{x_ny_n\}$ converges. On the other hand $\{(-1)^n\}$ diverges since given any $k \in \mathbb{N}$ we know that $|(-1)^k - (-1)^{k+1}| = 2$, so $\{(-1)^n\}$ is not a Cauchy sequence and therefore cannot converge.

□

Solution to Exercise 3.11. *Let r be a real number. There is a sequence of rational numbers converging to r.*

Proof. For each $n \in \mathbb{N}$ we can choose a rational number $q_n \in (r - \frac{1}{n}, r + \frac{1}{n})$ since we have shown there is a rational number between any two real numbers. By Theorem 3.1 and the sum rule for sequence limits we know that $\{r - \frac{1}{n}\} \to r$ and $\{r + \frac{1}{n}\} \to r$, so by the Squeeze Theorem it follows that $\{q_n\} \to r$.

□

Solution to Exercise 3.12. *Let $\{x_{n_i}\}$ be a subsequence of $\{x_n\}$. Then $n_i \geq i$ for each $i \in \mathbb{N}$.*

Proof. We know that $\{n_i\}$ is an increasing sequence of natural numbers by the definition of subsequence. Proceed by induction. First, $n_1 \in \mathbb{N}$ so $n_1 \geq 1$. Assume that $n_k \geq k$ for some natural number k. Then since $n_{k+1} > n_k \geq k$ and $k+1$ is the first natural number which exceeds k, it must follow that $n_{k+1} \geq k+1$. The result follows by induction.

□

Solution to Exercise 3.13. *If $|r| < 1$ then $\{r^n\} \to 0$.*

Proof. First, since $|r| < 1$ we know that $|r^{n+1}| = |r||r^n| < |r^n|$, so $\{|r^n|\}$ is a decreasing sequence which is bounded below by 0. Thus, from the Monotone Convergence Theorem we know that $\{|r^n|\}$ converges to its greatest lower bound L, and that $L \geq 0$ by the Comparison Theorem. We note that $\{|r^{n+1}|\}$ is a subsequence of $\{|r^n|\}$, and so by Theorem 3.11, we know that $\{|r^{n+1}|\} \to L$. However, $\{|r^{n+1}|\} = \{|r||r^n|\}$, so by the product rule for sequence limits, $\{|r^{n+1}|\} \to |r|L$. It follows that $L = |r|L$. Hence, either $r = 0$ or $L = 0$. If $r = 0$ then $\{r^n\} = \{0\} \to 0$. Otherwise $L = 0$, so $\{|r^n|\} \to 0$. From an Theorem 1.16 we know that $-|r^n| \leq r^n \leq |r^n|$, so by the Squeeze Theorem it follows that $\{r^n\} \to 0$.

□

Solution to Exercise 3.14. *Let p be a limit point of a set A and let $A \subseteq B$. Then p is a limit point of B.*

Proof. Let $\epsilon > 0$. Then there is a point $q \in (p-\epsilon, p+\epsilon) \cap A \setminus \{p\}$. Since $A \subseteq B$ we know that $q \in B$. Thus, p is a limit point of B. □

Solution to Exercise 3.15. *Let p be a limit point of $A \cup B$. Then either p is a limit point of A or p is a limit point of B.*

Proof. Assume that p is not a limit point of A. Then there is an $\epsilon_1 > 0$ so that $(p-\epsilon_1, p+\epsilon_1)$ contains no points of A other than p. This means that for all $\epsilon < \epsilon_1$ it is true that $(p-\epsilon, p+\epsilon)$ contains no point of A distinct from p. Since $(p-\epsilon, p+\epsilon)$ contains a point of $A \cup B$ distinct from p, it must follow that $(p-\epsilon, p+\epsilon)$ contains a point of B distinct from p. Thus, for every $\gamma > 0$ there is an $\epsilon < \min\{\epsilon_1, \gamma\}$ and there is a point $q \in (p-\epsilon, p+\epsilon) \cap B \setminus \{p\} \subseteq (p-\gamma, p+\gamma) \cap B \setminus \{p\}$, so p is a limit point of B. Hence, either p is a limit point of A or p is a limit point of B. □

Solution to Exercise 3.16. *Let E be a set and let $A \subseteq E$. Then $E \setminus A$ is open in E if and only if A is closed in E.*

Proof. Let $E \setminus A$ be open in E. Then there is an open set V so that $V \cap E = E \setminus A$. Since $\mathbb{R} \setminus V$ is closed, it follows that $(\mathbb{R} \setminus V) \cap E = A$ is closed in E.

Let A be closed in E. Then there is a closed set K so that $K \cap E = A$. Since $\mathbb{R} \setminus K$ is open, it follows that $(\mathbb{R} \setminus K) \cap E = E \setminus A$ is open in E. □

Solution to Exercise 3.17. *Let A_i be closed, non-empty and bounded for each natural number i, so that $A_1 \supseteq A_2 \supseteq A_3 \ldots$. Then $\bigcap_{i=1}^{\infty} A_i$ is non-empty.*

Proof. Choose $x_n \in A_n$ for each $n \in \mathbb{N}$. Then $\{x_n\} \subseteq A_1$, which is bounded, so $\{x_n\}$ is bounded. Hence, $\{x_n\}$ has a convergent subsequence $\{x_{n_i}\} \to p$. This means, for each $k \in \mathbb{N}$, the subsequence $\{x_{n_{i+k}}\} \subseteq A_k$ by Theorem 3.12, so $\{x_{n_{i+k}}\} \to p$ by Theorem 3.11 Since each A_k is closed we know that $p \in A_k$ for all $k \in \mathbb{N}$ by Theorem 3.16, and thus $p \in \bigcap_{i=1}^{\infty} A_i$. □

Solution to Exercise 3.18. *Let S be a bounded infinite set. Then S has a limit point.*

Proof. Let S be infinite. Choose $x_1 \in S$. If we have chosen x_i for $i \leq k$ then choose $x_{k+1} \in S \setminus \{x_1, x_2, x_3, ..., x_k\}$. Such a choice is always possible since S is infinite. Thus $\{x_n\}$ is a sequence of points of S, and $x_i \neq x_j$ for all $i \neq j$. By the Bolzano-Weierstrass Theorem, $\{x_n\}$ has a convergent subsequence $\{x_{n_i}\} \to p$. Thus, for every $\epsilon > 0$, the interval $(p-\epsilon, p+\epsilon)$ contains infinitely many elements of $\{x_{n_i}\}$ by Exercise 3.6 (each of which are elements of S), so p is a limit point of S.

□

Solution to Exercise 3.19. *The sets $\mathbb{Z}, \mathbb{R}, \mathbb{Q}$ are infinite sets.*

Proof. None of these points have last points (or first points) so each of these sets is infinite by Theorem 3.30. □

Solution to Exercise 3.20. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\} \to L$ if and only if the sequence tail $\{a_{n+k}\} \to L$.*

Proof. If $\{a_n\} \to L$ then since $\{a_{n+k}\}$ is a subsequence of $\{a_n\}$ it follows that $\{a_{n+k}\} \to L$ by Theorem 3.11. Next, assume $\{a_{n+k}\} \to L$. Let $\epsilon > 0$. We can choose $N \in \mathbb{N}$ so that if $n \geq N$ then $|a_{n+k} - L| < \epsilon$. Thus, if $n \geq N + k$ it follows that $|a_n - L| = |a_{n-k+k} - L| < \epsilon$. Hence, $\{a_n\} \to L$. □

Solution to Exercise 3.21. *Let $\{a_n\}$ be a sequence and k be a natural number. Then $\{a_n\}$ is bounded above if and only if the sequence tail $\{a_{n+k}\}$ is bounded above, and $\{a_n\}$ is bounded below if and only if the sequence tail $\{a_{n+k}\}$ is bounded below.*

Proof. If there is some M so that $a_n \leq M$ for all $n \in \mathbb{N}$ then $a_n \leq M$ for all $n \geq k$, so if $\{a_n\}$ is bounded above then $\{a_{n+k}\}$ is bounded above.

If there is some M so that $a_n \leq M$ for all $n \geq k$ then $a_n \leq \max\{a_1, a_2, ..., a_{k-1}, M\}$ for all $n \in \mathbb{N}$, so $\{a_n\}$ is bounded above.

If there is some M so that $a_n \geq M$ for all $n \in \mathbb{N}$ then $a_n \geq M$ for all $n \geq k$, so if $\{a_n\}$ is bounded below then $\{a_{n+k}\}$ is bounded below.

If there is some M so that $a_n \geq M$ for all $n \geq k$ then $a_n \geq \min\{a_1, a_2, ..., a_{k-1}, M\}$ for all $n \in \mathbb{N}$, so $\{a_n\}$ is bounded below.

□

Chapter 4

Limits and Continuity

Definition 26

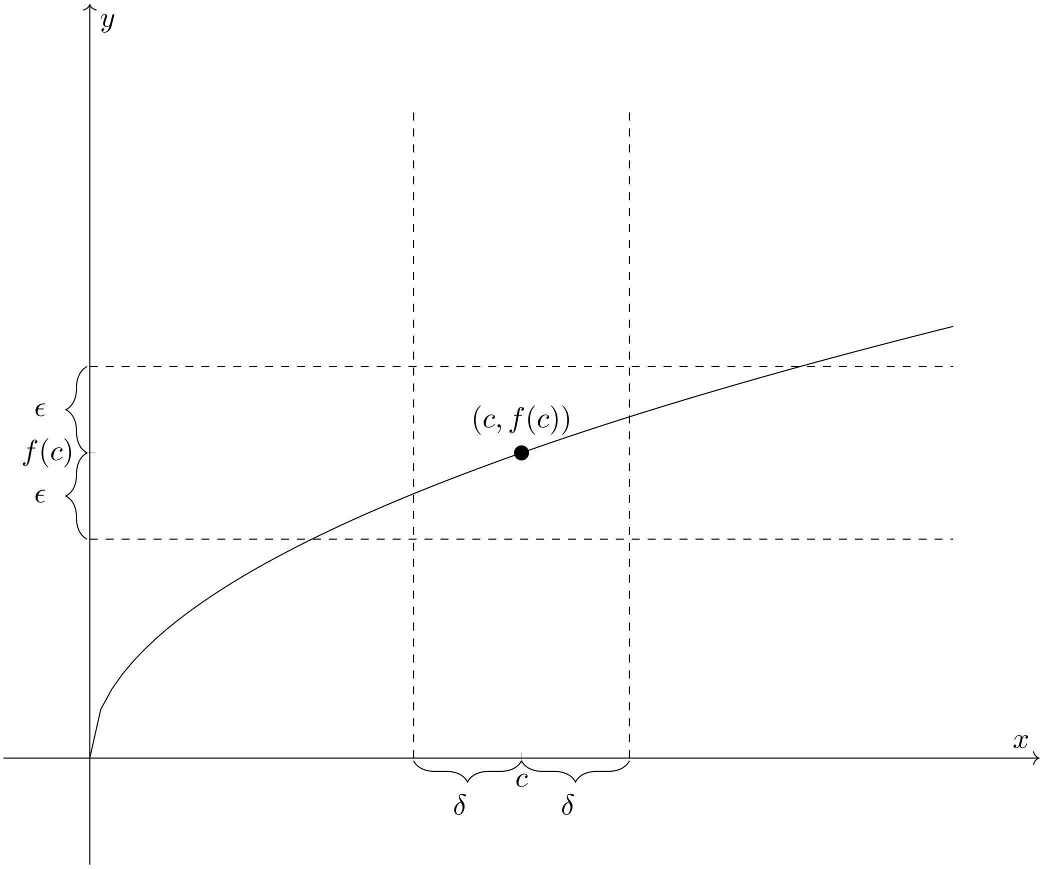
Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. Then we say that $\lim_{x \to c} f(x) = L$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - L| < \epsilon$.

We say that f is *continuous* at the point $z \in D$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $|x - z| < \delta$ and $x \in D$ then $|f(x) - f(z)| < \epsilon$. We say that a function f is continuous if it is continuous at every point in its domain. We say that f is continuous on the set E if f is continuous at every point of E.

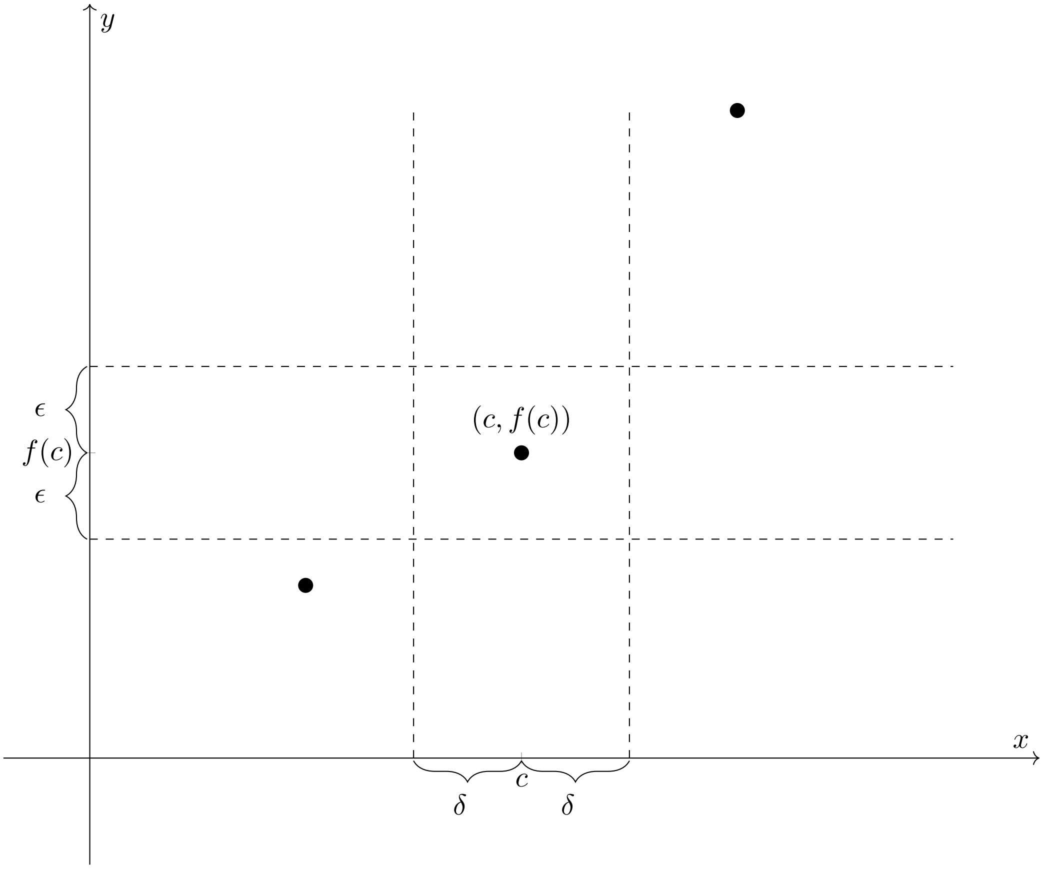
The preceding definition gives us another reason for using the term ”limit” point to describe a limit point. A point p is a limit point of set D if and only if there are functions with domain D that have a limit at the point p. People often think of continuity using the graph of a function as saying that a function is continuous if its graph has no breaks in it. That is an accurate definition (once the idea of having no breaks is formalized) if the domain of the function is an interval, but it is not correct in general. Using the definition above, we notice that there are some functions which are continuous that do not have connected graphs. For instance, if the domain of a function f is the integers then the function is always continuous, because for any integer k in the domain and $\epsilon > 0$, if $|x - k| < 1$ and $x \in \mathbb{Z}$ then $x = k$ so $|f(x) - f(k)| = 0 < \epsilon$. The graph of a function whose domain is the integers is just a discrete collection of points, but it is still a continuous function. Likewise, the function $f(x) = \dfrac{1}{x}$ is continuous even though its graph is broken into two pieces. The value zero is not in the domain of that function, and for every value in the domain of the function, $f(x)$ is continuous, making f a continuous function. A function can be continuous at a point of the domain where the limit does not exist. Specifically, functions are always continuous at points of their domains which are not limit points of the domain (and limits of those functions cannot exist as x approaches those points).

Graphically, as long as, given any ϵ-radius interval centered at L, there is a δ-radius interval centered at c so that portion of the graph of $y = f(x)$ whose x-coordinates lie within $(c - \delta, c + \delta)$ has y-coordinates within $(f(c) - \epsilon, f(c) + \epsilon)$, meaning that the section of the graph of f inside the vertical band between lines $x = c - \delta$ and $x = c + \delta$ also lies within the horizontal band between $y = f(c) - \epsilon$ and $y = f(c) + \epsilon$.

Continuity at c

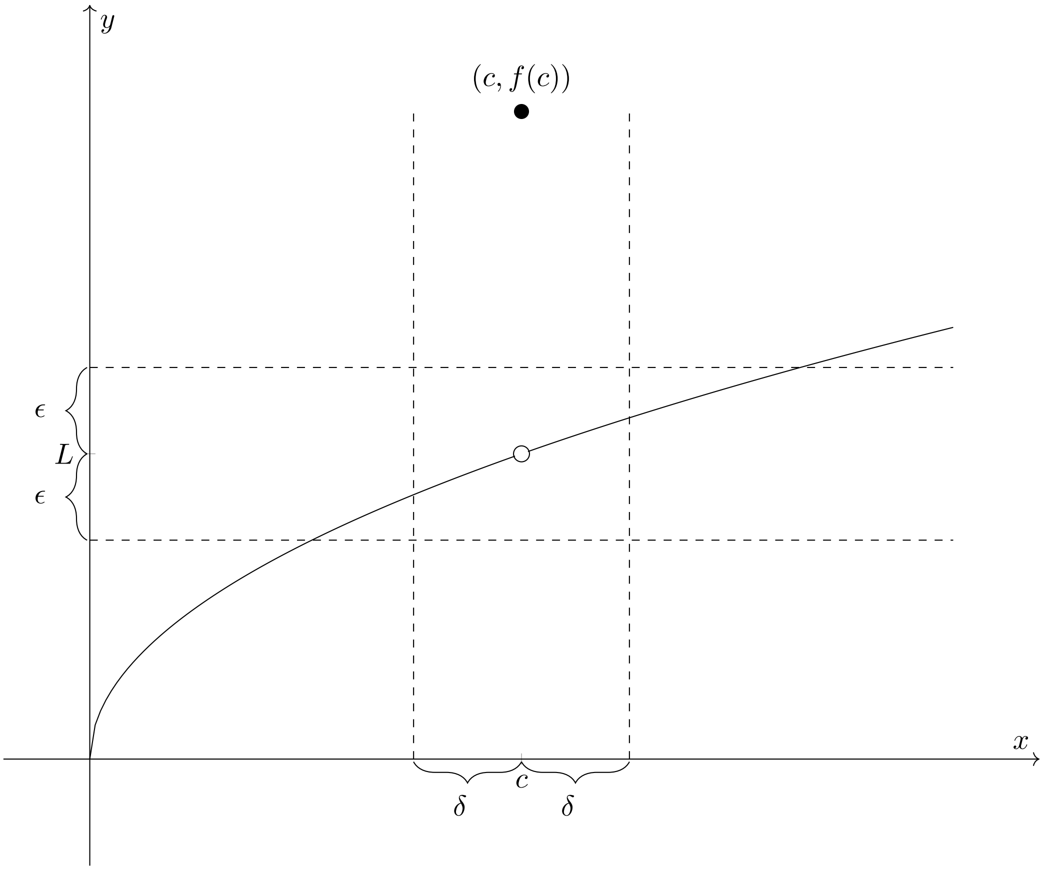


Continuity at Isolated Point c



The picture for $\lim\limits_{x \to c} f(x) = L$ is similar, except that c must be a limit point of the domain, not an isolated point, and it need not be in the domain. Furthermore, the value of $f(c)$ has no bearing on the limit if c is in the domain (only the points close to but not equal to c are relevant to the definition of function limit).

Limit at c



We note, at this point, that it is instructive to look at trigonometric functions from time to time, but we are not developing trigonometry rigorously, and are assuming that properties of geometry and trigonometry were proven in another text and are now being assumed to be true.

Theorem 4.1. *The Sequential Characterization of Limits. Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. Then $\lim\limits_{x \to c} f(x) = L$ if and only if for every sequence $\{x_n\} \subseteq D \setminus \{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$.*

Proof. First, assume that $\lim\limits_{x \to c} f(x) = L$. Let $\{x_n\} \subseteq D \setminus \{c\}$ such that $\{x_n\} \to c$. Let $\epsilon > 0$. Then for some $\delta > 0$, we know that if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - L| < \epsilon$. Since $\{x_n\} \to c$, we can find $N \in \mathbb{N}$ so that if $n \geq N$ then $|x_n - c| < \delta$, and since $\{x_n\} \subseteq D \setminus \{c\}$ it follows that if $n \geq N$ then $0 < |x_n - c| < \delta$. Hence, if $n \geq N$ then $|f(x_n) - L| < \epsilon$, so $\{f(x_n)\} \to L$.

Next, assume that for every sequence $\{x_n\} \subseteq D \setminus \{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$. Suppose that $\lim\limits_{x \to c} f(x) \neq L$. Then we can find an $\epsilon > 0$ so that for every $\delta > 0$ there is some $x \in D \setminus \{c\}$ so that $|x - c| < \delta$ but $|f(x) - L| \geq \epsilon$. For each $n \in \mathbb{N}$ we choose $x_n \in D \setminus \{c\}$ so that $|x_n - c| < \dfrac{1}{n}$ and $|f(x_n) - L| \geq \epsilon$. Since $c - \dfrac{1}{n} \leq x_n \leq c + \dfrac{1}{n}$, we know by the Squeeze Theorem that $\{x_n\} \to c$. But $\{f(x_n)\} \not\to L$, contradicting our assumption.

□

Theorem 4.2. *The Sequential Characterization of Continuity. Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and $c \in D$. Then f is continuous at c if and only if for every sequence $\{x_n\} \subseteq D$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to f(c)$.*

The proof is similar to that of the preceding theorem and is left as an exercise.

Theorem 4.3. *The functions $f(x) = k$ and $g(x) = x$ on domain D are continuous.*

Proof. Let $c \in D$, and let $\{x_n\} \to c$ for some $\{x_n\} \subseteq D$. By Theorem 3.2 we know $\{f(x_n)\} = \{k\} \to k = f(c)$. Also, we know that $\{g(x_n)\} = \{x_n\} \to c = g(c)$. Thus, f and g are continuous. □

Theorem 4.4. *Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and $c \in D$.*

(a) Let c be a limit point of D. Then f is continuous at c if and only if $\lim\limits_{x\to c} f(x) = f(c)$.

(b) If c is not a limit point of D then f is continuous at c.

Proof. (a) First, assume that f is continuous at c. Let $\epsilon > 0$. We know that for some $\delta > 0$, if $|x - c| < \delta$ and $x \in D$ then $|f(x) - f(c)| < \epsilon$. Hence, if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - f(c)| < \epsilon$, so $\lim\limits_{x\to c} f(x) = f(c)$.

Next, assume that $\lim\limits_{x\to c} f(x) = f(c)$. Let $\epsilon > 0$. We know that for some $\delta > 0$, if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - f(c)| < \epsilon$, but if $x = c$ then $|f(x) - f(c)| = 0 < \epsilon$ as well. Hence, if $|x - c| < \delta$ and $x \in D$ then $|f(x) - f(c)| < \epsilon$, so f is continuous at c.

(b) Since c is an isolated point of D, we can find $\delta > 0$ so that the only point of D whose distance from c is less than δ is c. Hence, if $|x - c| < \delta$ and $x \in D$ then $x = c$, so $|f(x) - f(c)| = 0$ which is less than any positive number ϵ and therefore f is continuous at c.

□

Theorem 4.5. *(a) Let $f : dom(f) \to \mathbb{R}$ be continuous at c and let $f(c) \neq 0$. Then there is some $\delta > 0$ so that $|f(x)| > \dfrac{|f(c)|}{2}$ if $x \in (c - \delta, c + \delta) \cap dom(f)$.*

(b) If f is continuous at $c \in (a, b) \subseteq dom(f)$ and $f(c) \neq 0$ then there is some $\delta > 0$ so that $(c - \delta, c + \delta) \subseteq dom(f)$.

(c) Let $g : dom(g) \to \mathbb{R}$. If $\lim\limits_{x\to c} g(x) = L \neq 0$ then there is some $\delta > 0$ so that $|g(x)| > \dfrac{|g(c)|}{2}$ if $x \in (c - \delta, c + \delta) \cap dom(g) \setminus \{c\}$ and c is a limit point of $dom(\dfrac{1}{g})$.

Proof. (a) Choose $\delta > 0$ so that if $|x - c| < \delta$ and $x \in dom(f)$ then $|f(x) - f(c)| < \dfrac{|f(c)|}{2}$. It follows that if $x \in (c - \delta, c + \delta) \cap dom(f)$ then $f(x) \geq \dfrac{|f(c)|}{2}$ by the Triangle Inequality, so $\dfrac{1}{f(x)}$ is defined. If $c \in (a,b) \subseteq dom(f)$ then set $\delta_1 = \min\{\delta, |c-a|, |c-b|\}$ then. If $x \in (c-\delta_1, c+\delta_1)$ then $x \in dom(f)$ and $f(x) \neq 0$ and therefore $x \in dom(\frac{1}{f})$, so $(c-\delta_1, c+\delta_1) \subseteq dom(\frac{1}{f})$.

(b) Let $\lim\limits_{x\to c} g(x) = L \neq 0$. Choose $\delta > 0$ so that if $0 < |x-c| < \delta$ and $x \in dom(g)$ then $|g(x) - L| < \dfrac{|L|}{2}$, so $|g(x)| > \dfrac{|L|}{2}$. It follows that if $x \in (c-\delta, c+\delta) \cap dom(g) \setminus \{c\}$ then $\dfrac{1}{g(x)}$ is defined. Let $\epsilon > 0$. Let $\gamma = \min\{\delta, \epsilon\}$. Since c is a limit point of $dom(g)$ we can find a point $q \neq 0$ so that $q \in (c-\gamma, c+\gamma) \cap dom(g)$, which means that $q \in dom(\frac{1}{g}) \cap (c-\epsilon, c+\epsilon)$ and therefore c is a limit point of $\dfrac{1}{g}$.

□

Example 4.1. *(a) Let $f(x) = x$ if $x \neq 2$ and let $f(2) = 5$. Find $\lim\limits_{x\to 2} f(x)$.*

(b) Let $f(x) = 0$ if $x \in \mathbb{Q}$ and let $f(x) = 1$ if $x \in \mathbb{R} \setminus \mathbb{Q}$. Prove that f is discontinuous at every real number.

(c) Let $f : \mathbb{N} \to \mathbb{N}$ be defined by $f(x) = x_n$ for some sequence $\{x_n\}$. At what points is f continuous?

Solution. (a) $\lim\limits_{x\to 2} f(x) = 2$ since $g(x) = x$ is continuous by Exercise 4.3 (the value of the function at the point approached does not affect the limit).

(b) Let $r \in \mathbb{R}$ and let $\delta > 0$. Then $(p-\delta, p+\delta)$ contains both an irrational number α and a rational number q. Thus, $|f(\alpha) - f(q)| = 1$, so if $|f(\alpha) - f(p)| < \dfrac{1}{2}$ then $|f(p) - f(\alpha)| > \dfrac{1}{2}$, and if $|f(\alpha) - f(\alpha)| < \dfrac{1}{2}$ then $|f(p) - f(q)| > \dfrac{1}{2}$. Hence, there is no $\delta > 0$ so that for every x so that $|x - p| < \delta$ it is true that $|f(x) - f(p)| < \dfrac{1}{2}$. Hence, f is not continuous at any point p.

(c) f is continuous at every point in its domain. To see this, let $\epsilon > 0$. For any $m \in \mathbb{Z}$, if $|x - m| < 1$ and $x \in dom(f)$ then $x = m$ which means $|f(x) - m| = 0 < \epsilon$.

□

Theorem 4.6. *Let $f, g : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. Let $\lim\limits_{x\to c} f(x) = r$ and $\lim\limits_{x\to c} g(x) = s$. Then the following are true:*

(a) Sum Rule (for limits): $\lim\limits_{x\to c} af(x) + bg(x) = ar + bs$

(b) Product Rule (for limits): $\lim\limits_{x\to c} f(x)g(x) = rs$

(c) Quotient Rule (for limits): If $s \neq 0$ then $\lim\limits_{x\to c} \dfrac{f(x)}{g(x)} = \dfrac{r}{s}$

Proof. Let $\{x_n\} \subseteq D \setminus \{c\}$ so that $\{x_n\} \to c$. Then by Theorem 4.1 we know that $\{f(x_n)\} \to r$ and $\{g(x_n)\} \to s$, so by Theorem 3.9, $\{af(x_n) + bg(x_n)\} \to ar + bs$, $\{f(x_n)g(x_n)\} \to rs$, and if $\{x_n\} \subseteq dom(\frac{f(x)}{g(x)}) \setminus \{c\}$ then $\{\frac{f(x_n)}{g(x_n)}\} \to \frac{r}{s}$ (and if $s \neq 0$ then by Theorem 4.5, we know that c is a limit point of $dom(\frac{f}{g})$). Hence, by the Theorem 4.1, the result follows.

□

Note that each of these can be proven directly without using the Sequential Characterization of Limits. It is to our advantage to develop sequences for other proofs as well, and having developed sequences it is arguably a waste of time to duplicate everything for function limits.

We may not always refer to Theorems 4.1 and 4.2 when we use them. Readers are encouraged to think of the sequential methods of characterizing limits and continuity as being more like a second definition than a theorem. The sum rule is usually written without the constants a and b multiplied by the functions, but is is convenient for us to include those constants.

Theorem 4.7. *Let f, g be continuous at c. Then $f + g$, fg are continuous at c and $\frac{f}{g}$ is continuous at c if $g(c) \neq 0$.*

Proof. Let $\{x_n\} \subseteq D$ so that $\{x_n\} \to c$. Then by Theorem 4.2 we know that $\{f(x_n)\} \to f(c)$ and $\{g(x_n)\} \to g(c)$. By Theorem 3.9, it follows that $\{f(x_n) + g(x_n)\} \to f(c) + g(c)$, $\{f(x_n)g(x_n)\} \to f(c)g(c)$, and if $\{x_n\} \subseteq dom(\frac{f(x)}{g(x)})$ then $\{\frac{f(x_n)}{g(x_n)}\} \to \frac{f(c)}{g(c)}$ if $g(c) \neq 0$. Thus, the result follows from Theorem 4.2.

□

Theorem 4.8. *Squeeze Theorem (for limits). Let $f, g, h : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. If there is a $\delta_1 > 0$ so that $f(x) \leq g(x) \leq h(x)$ or $h(x) \leq g(x) \leq f(x)$ for all $x \in D$ so that $0 < |x - c| < \delta_1$ and $\lim_{x \to c} f(x) = L = \lim_{x \to c} h(x)$ then $\lim_{x \to c} g(x) = L$.*

Proof. Choose $\delta_2 > 0$ so that if $0 < |x - c| < \delta_2$ then $|f(x) - L| < \epsilon$. Choose $\delta_3 > 0$ so that if $0 < |x - c| < \delta_3$ then $|h(x) - L| < \epsilon$. Let $\delta = \min(\delta_1, \delta_2, \delta_3)$. If $0 < |x - c| < \delta$ then $L - \epsilon < f(x) \leq g(x) \leq h(x) < L + \epsilon$ or $L - \epsilon < h(x) \leq g(x) \leq f(x) < L + \epsilon$, so $|g(x) - L| < \epsilon$, so $\lim_{x \to c} g(x) = L$.

□

As with sequences, the Squeeze Theorem for limits helps us evaluate many function limits.

Example 4.2. *Prove that $\lim_{x \to 0} x \sin(\frac{1}{x}) = 0$.*

Solution. We assume it is known that $-1 \leq \sin(\frac{1}{x}) \leq 1$ (we are assuming properties of trigonometric functions which are not the result of limits are known in this text). In that case $x\sin(\frac{1}{x})$ is in the interval $[-x,x]$ for all $x \neq 0$. If $x > 0$ then $-x \leq x\sin(\frac{1}{x}) \leq x$, and if $x < 0$ then $x \leq x\sin(\frac{1}{x}) \leq -x$. Since $\lim_{x\to 0} x = 0 = \lim_{x\to 0} -x$ we conclude that $\lim_{x\to 0} x\sin(\frac{1}{x}) = 0$ by the Squeeze Theorem. □

Theorem 4.9. *Comparison Theorem (for function limits). Let $f, g : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. If there is a $\delta > 0$ so that $f(x) \leq g(x)$ for all $x \in D$ so that $0 < |x - c| < \delta$ and $\lim_{x\to c} f(x) = s$ and $\lim_{x\to c} g(x) = t$ then $s \leq t$.*

Proof. Let $\{x_n\} \subseteq D\backslash\{c\}$ so that $\{x_n\} \to c$. Then by Theorem 4.1 we know that $\{f(x_n)\} \to s$ and $\{g(x_n)\} \to t$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $0 < |x_n - c| < \delta$. If $n \geq k$ then $f(x_n) \leq g(x_n)$ so the hypotheses of the Comparison Theorem for sequences are satisfied, and hence $s \leq t$.

□

Note that when we refer to the Comparison or Squeeze theorems we usually assume it is clear from context which form (sequences or limits) is being cited, so we typically do not say "Squeeze Theorem for Limits" in arguments and instead just say "Squeeze Theorem."

Theorem 4.10. *Let $f, g : D \to L$ and let c be a limit point of D. Let $\lim_{x\to c} f(x) = L$.*

(a) If $\lim_{x\to c} f(x) + g(x) = L + R$ then $\lim_{x\to c} g(x) = R$.

(b) If $L \neq 0$ and $\lim_{x\to c} f(x)g(x) = LR$ then $\lim_{x\to c} g(x) = R$

Proof. Let $\{x_n\} \subseteq D\backslash\{c\}$ so that $\{x_n\} \to c$. Then by Theorem 4.1 we know that $\{f(x_n)\} \to L$. Thus, by Exercises 3.8 and 3.9 we can conclude that $\{g(x_n)\} \to R$ in (a) and (b) respectively, which means that $\lim_{x\to c} g(x) = R$.

□

Theorem 4.11. *If c is a limit point of $dom(f \circ g)$ and $\lim_{x\to c} g(x) = L$ and $f(x)$ is continuous at L then $\lim_{x\to c} f(g(x)) = f(L)$. If L is a limit point of the domain of f then $\lim_{x\to c} f(g(x)) = \lim_{y\to L} f(y)$.*

Proof. Let $\{x_n\} \subseteq dom(f \circ g)\backslash\{c\}$ so that $\{x_n\} \to c$. Then by The Sequential Characterization of Limits we know that $\{g(x_n)\} \to L$ and since f is continuous at L we know that $\{f(g(x_n))\} \to f(L)$ by The Sequential Characterization of Continuity. Thus, by The Sequential Characterization of Limits we know that $\lim_{x\to c} f(g(x)) = f(L)$. If L is a limit point of the domain of $f(x)$ then by Theorem 4.4, $\lim_{y\to L} f(x) = f(L) = \lim_{x\to c} f(g(x))$.

□

Theorem 4.12. *Let f be continuous at $g(c)$ and g be continuous at c. Then $f \circ g$ is continuous at c.*

Proof. Let $\{x_n\} \subseteq dom(f \circ g)$ so that $\{x_n\} \to c$. Then by The Sequential Characterization of Continuity we know that $\{g(x_n)\} \to g(c)$ and hence $\{f(g(x_n))\} \to f(g(c))$, so $f \circ g$ is continuous at c.

□

There are advantages and disadvantages to proving theorems about infinite limits in a brief development of advanced calculus. For the most part, we will not need theorems on infinite limits, but we will develop them in more detail in the section on infinite limits in the Supplementary Materials section for those who are interested. A problem with these sorts of definitions is that the more general forms of theorems involving limits that could be infinite at points or possibly at infinity or negative infinity tend to take a fairly simple idea for a proof and require repetitions for many cases to make it rigorous. However, it is worthwhile to understand what these definitions are whether we spend a lot of time considering theorems about them or not.

Definition 27

Let $f : D \to \mathbb{R}$, where D is not bounded above. We say that $\lim\limits_{x\to\infty} f(x) = L$ if, for every $\epsilon > 0$, there is an M so that if $x > M$ and $x \in D$ then $|f(x) - L| < \epsilon$. We say that $\lim\limits_{x\to\infty} f(x) = \infty$ if, for every T, there is an M so that if $x > M$ and $x \in D$ then $f(x) > T$. We say that $\lim\limits_{x\to\infty} f(x) = -\infty$ if, for every T, there is an M so that if $x > M$ and $x \in D$ then $f(x) < T$.
Let $f : D \to \mathbb{R}$, where D is not bounded below. We say that $\lim\limits_{x\to-\infty} f(x) = L$ if, for every $\epsilon > 0$, there is an M so that if $x < M$ and $x \in D$ then $|f(x) - L| < \epsilon$. We say that $\lim\limits_{x\to-\infty} f(x) = \infty$ if, for every T, there is an M so that if $x < M$ and $x \in D$ then $f(x) > T$. We say that $\lim\limits_{x\to-\infty} f(x) = -\infty$ if, for every T, there is an M so that if $x < M$ and $x \in D$ then $f(x) < T$.
Let $f : D \to \mathbb{R}$, where c is a limit point of D. We say that $\lim\limits_{x\to c} f(x) = \infty$ if, for every $M \in \mathbb{R}$, there is a $\delta > 0$ so that if $0 < |x - c| < \delta$ and $x \in D$ then $f(x) > M$. Similarly, we define $\lim\limits_{x\to c} f(x) = -\infty$ if, for every $M \in \mathbb{R}$, there is a $\delta > 0$ so that if $0 < |x - c| < \delta$ and $x \in D$ then $f(x) < M$.

Note that if $D = \mathbb{N}$, where $f(n) = x_n$ for all $n \in \mathbb{N}$, then $\lim\limits_{n\to\infty} f(n) = \lim\limits_{n\to\infty} x_n = L$ is equivalent to $\{x_n\} \to L$. The details of this are left as an exercise.

Theorem 4.13. *Sequential Characterization of Limits for Infinite Limits at real numbers.*
Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D. Then $\lim\limits_{x\to c} f(x) = \infty$ (or $-\infty$ respectively) if and only if for every sequence $\{x_n\} \subseteq D\backslash\{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to \infty$ (or $-\infty$ respectively).

Proof. First, assume that $\lim\limits_{x\to c} f(x) = \infty$. Then let $M \in \mathbb{R}$. We can choose $\delta > 0$ so that if $x \in D$ and $0 < |x - c| < \delta$ then $f(x) > M$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - c| < \delta$.

Then since, $\{x_n\} \subseteq D \setminus \{c\}$ it follows that if $n \geq k$ then $0 < |x_n - c| < \delta$ and $f(x_n) > M$. Thus, $\{f(x_n)\} \to \infty$.

Likewise, if $\lim\limits_{x \to c} f(x) = -\infty$ and $M \in \mathbb{R}$ then we can choose $\delta > 0$ so that if $x \in D$ and $0 < |x - c| < \delta$ then $f(x) < M$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - c| < \delta$. Then since, $\{x_n\} \subseteq D \setminus \{c\}$ it follows that if $n \geq k$ then $0 < |x_n - c| < \delta$ and $f(x_n) < M$. Thus, $\{f(x_n)\} \to -\infty$.

Next, assume that for every sequence $\{x_n\} \subseteq D \setminus \{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to \infty$ (or $-\infty$ respectively). Suppose that $\lim\limits_{x \to c} f(x) \neq \infty$ (or $-\infty$ respectively). Then for some $M \in \mathbb{R}$, for every $\delta > 0$ we can choose $x \in D$ so that $0 < |x - c| < \delta$ but $f(x) \leq M$ (or $f(x) \geq M$ respectively). Thus, for every integer $n \in \mathbb{N}$ we can pick $x_n \in D \setminus \{c\}$ so that $|x_n - c| < \dfrac{1}{n}$ and $f(x_n) \leq M$ (or $f(x_n) \geq M$ respectively). Hence, $\{x_n\} \to c$ but $\{f(x_n)\} \not\to \infty$ (or $-\infty$ respectively), a contradiction. □

Theorem 4.14. *Let $f : D \to \mathbb{R}$ be a function so $(a, \infty) \subset D$ for some $a \in \mathbb{R}$. Then $\lim\limits_{x \to \infty} f(x) = L$ if and only if $\lim\limits_{x \to 0^+} f(\dfrac{1}{x}) = L$. Likewise, if $g : D \to \mathbb{R}$ is a function so $(-\infty, a) \subseteq D$ then $\lim\limits_{x \to -\infty} g(x) = L$ if and only if $\lim\limits_{x \to 0^-} g(\dfrac{1}{x}) = L$.*

Proof. Let $\epsilon > 0$. Assume $\lim\limits_{x \to \infty} f(x) = L$. Choose $M > \max(a, 0)$ so that if $x > M$ then $|f(x) - L| < \epsilon$. Then if $0 < x < \dfrac{1}{M}$ it follows that $\dfrac{1}{x} > M$ so $|f(\dfrac{1}{x}) - L| < \epsilon$. Hence, $\lim\limits_{x \to 0^+} f(\dfrac{1}{x}) = L$. Similarly, if $\lim\limits_{x \to 0^+} f(\dfrac{1}{x}) = L$ then we can choose $\delta > 0$ so that if $0 < x < \delta$ then $|f(\dfrac{1}{x}) - L| < \epsilon$. Thus, if $M > \max(a, \dfrac{1}{\delta})$ then if $y > M$ we know that $y \in (a, \infty)$ and $0 < \dfrac{1}{y} < \delta$, so $f(y) = f(\dfrac{1}{x})$ for some $x = \dfrac{1}{y}$ so that $0 < x < \delta$ which means that $|f(y) - L| < \epsilon$. Hence, $\lim\limits_{x \to \infty} f(x) = L$.

The proof that if $g : (-\infty, a) \to \mathbb{R}$ then $\lim\limits_{x \to -\infty} g(x) = L$ if and only if $\lim\limits_{x \to 0^-} g(\dfrac{1}{x}) = L$ is similar. Assume $\lim\limits_{x \to -\infty} g(x) = L$. Choose $M < \min(a, 0)$ so that if $x < M$ then $|g(x) - L| < \epsilon$. Then if $\dfrac{1}{M} < x < 0$ it follows that $\dfrac{1}{x} < M$ so $|g(\dfrac{1}{x}) - L| < \epsilon$. Hence, $\lim\limits_{x \to 0^-} g(\dfrac{1}{x}) = L$. Similarly, if $\lim\limits_{x \to 0^-} g(\dfrac{1}{x}) = L$ then we can choose $\delta > 0$ so that if $-\delta < x < 0$ then $|g(\dfrac{1}{x}) - L| < \epsilon$. Thus, if $M < \min(a, \dfrac{-1}{\delta})$ then if $y < M$ we know that $y \in (-\infty, a)$ and $-\delta < \dfrac{1}{y} < 0$, so $g(y) = g(\dfrac{1}{x})$ for some $x = \dfrac{1}{y}$ so that $-\delta < x < 0$ which means that $|g(y) - L| < \epsilon$. Hence, $\lim\limits_{x \to -\infty} g(x) = L$. □

There are other results about infinite limits that we do not need for this development which are addressed in the Supplementary Materials section.

Sometimes it is helpful to specifically look at the portion of the domain of a function which is greater than or less than a point, and the corresponding one sided limit at that point obtained from such a restriction.

Definition 28

Let $f : D \to \mathbb{R}$, and let $A \subset D$. We define the *restriction of f to A* to be the function $g : A \to \mathbb{R}$ defined by setting $g(x) = f(x)$ for all $x \in A$. We use the notation $f|_A$ to denote this restriction. We define $\lim\limits_{x\to c^+} f(x) = \lim\limits_{x\to c} f|_{D\cap(c,\infty)}$ and $\lim\limits_{x\to c^-} f(x) = \lim\limits_{x\to c} f|_{D\cap(-\infty,c)}$ if these exist (this definition holds for both finite and infinite limits).

Another way of stating those two definitions (for finite limits) is that if c is a limit point of the points of D preceding c then we define $\lim\limits_{x\to c^-} f(x) = L$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $c - \delta < x < c$ and $x \in D$ then $|f(x) - L| < \epsilon$, and if c is a limit point of the points of D greater than c then $\lim\limits_{x\to c^+} f(x) = L$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $c < x < c + \delta$ and $x \in D$ then $|f(x) - L| < \epsilon$. These limits are called one sided limits, or limits approaching from below or above (or from the left or right).

Theorem 4.15. *Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$ and c is a limit point of D, $D \cap (c, \infty)$ and $D \cap (-\infty, c)$. Then $\lim\limits_{x\to c} f(x) = L$ if and only if $\lim\limits_{x\to c^-} f(x) = L$ and $\lim\limits_{x\to c^+} f(x) = L$.*

Proof. First, assume that $\lim\limits_{x\to c} f(x) = L$, and pick $\delta > 0$ so that if $0 < |x - c| < \delta$ then $|f(x) - L| < \epsilon$. Then if $c - \delta < x < c$ or $c < x < c + \delta$ it follows that $|f(x) - L| < \epsilon$. Hence, $\lim\limits_{x\to c^-} f(x) = L$ and $\lim\limits_{x\to c^+} f(x) = L$.

Next, assume that $\lim\limits_{x\to c^-} f(x) = L$ and $\lim\limits_{x\to c^+} f(x) = L$. Choose $\delta_1, \delta_2 > 0$ so that if $c - \delta_1 < x < c$ then $|f(x) - L| < \epsilon$ and if $c < x < c + \delta_2$ then $|f(x) - L| < \epsilon$. Let $\delta = \min(\delta_1, \delta_2)$. Then if $0 < |x - c| < \delta$, it must follow that either $c - \delta_1 < x < c$ or $c < x < c + \delta_2$, so $|f(x) - L| < \epsilon$. Hence, $\lim\limits_{x\to c} f(x) = L$.

□

By adjusting the domain in most of these limit theorems, we can see that most of the theorems we have proven are also proven for one sided limits (just restrict our hypothesis to domains containing points on only one side of c).

If readers are sufficiently comfortable with more abstract topological ideas then it may be helpful to introduce the definition of compactness now. However, the results about compact and connected sets can also be ignored until later without losing much. We include these theorems in the Supplementary Materials section so that readers who are interested can look at the topological theorems first. It is almost certainly a good idea to learn these theorems eventually, but it is possible to wait until discussing more abstract ideas or even until generalizing theorems to $\mathbb{R}^n$ if that is preferable. If the reader chooses to study the topological theorems now then some of the proofs of theorems in the remainder of this section can be abbreviated once these topological foundations are established. These alternate proofs are in the Supplementary Materials chapter in the "Topology of the Real Line" section.

Theorem 4.16. *The Extreme Value Theorem. Let $f : K \to \mathbb{R}$ be continuous, where K is closed and bounded. Then there are points $s, t \in K$ so $f(s) \leq f(x) \leq f(t)$ for every $x \in K$.*

Proof. Part 1: We wish to show that f is bounded. Suppose f is not bounded. Then for every $n \in \mathbb{N}$ we may choose $x_n \in K$ so that $|f(x_n)| > n$. Since $\{x_n\}$ is bounded, by the Bolzano-Weierstrass Theorem we can find a convergent subsequence $\{x_{n_k}\} \to p$, where $p \in K$ because K is closed. Since f is continuous it follows that $\{f(x_{n_k})\} \to f(p)$. But for every $M > 0$ we can find $N \in \mathbb{N}$ so that $N > M$ and thus $|f(x_{n_N})| > n_N \geq N > M$, and hence $\{f(x_{n_k})\}$ is not bounded and therefore cannot converge. This contradiction implies that f is bounded.

Part 2: We wish to show that there is a point $t \in K$ so that $f(t) = \sup(f(K))$. Let $l = \sup(f(K))$. By the Approximation Property, for every $n \in \mathbb{N}$ we can choose a point $z_n \in K$ so that $l - \dfrac{1}{n} < f(z_n) \leq l$. Since $\{z_n\}$ is bounded, by the Bolzano-Weierstrass Theorem we can find a convergent subsequence $\{z_{n_k}\} \to t \in K$ (where $t \in K$ because K is closed). Since f is continuous, we know that $\{f(z_{n_k})\} \to f(t)$, and by the Squeeze Theorem $\{f(z_{n_k})\} \to l$. Hence, $f(t) = l$.

Finally, by Part 2 we may find $s \in K$ so that $-f(s)$ is the maximum of $-f(K)$ and thus $f(s) \leq f(x)$ for all $x \in K$.

□

Theorem 4.17. *Intermediate Value Theorem. Let $f : [a, b] \to \mathbb{R}$ be continuous and let r be between $f(a)$ and $f(b)$. Then $f(c) = r$ for some $c \in (a, b)$.*

Proof. First, assume $f(a) < r < f(b)$. Let $S = \{x \in [a, b] | f(x) < r\}$. Note that $a \in S$ and b is an upper bound for S, so S has a least upper bound c. Thus, for each $n \in \mathbb{N}$ we can choose $x_n \in (c - \dfrac{1}{n}, c]$ so that $x_n \in S$. By the Squeeze Theorem, $\{x_n\} \to c$, so by the Comparison Theorem $\{f(x_n)\} \to f(c) \leq r$. Hence, $c < b$. For $x \in (c, b)$ we know that $f(x) \geq r$, so by the Comparison Theorem we know that $\lim\limits_{x \to c^+} f(x) = f(c) \geq r$. Thus, $f(c) = r$.

Note that if $f(b) < r < f(a)$ then $-f(a) < -r < -f(b)$ so for some $c \in (a, b)$ we know that $-f(c) = -r$ and therefore $f(c) = r$.

□

The following is another proof. It is a little longer, but it is easier to see pictorially.

Proof. Assume that $f(a) < r < f(b)$. Let $a_1 = a$ and $b_1 = b$. If $f(\dfrac{a_1 + b_1}{2}) \geq r$ then set $a_1 = a_2$ and $b_2 = \dfrac{a_1 + b_1}{2}$. Otherwise, set $a_2 = \dfrac{a_1 + b_1}{2}$ and $b_1 = b_2$. Proceeding inductively, if we have chosen a_i, b_i for $1 \leq i \leq k$ so that $a_1 \leq a_2 \leq ... \leq a_k \leq b_k \leq b_{k-1} \leq ... \leq b_1$,

$f(a_i) \leq r \leq f(b_i)$ and $b_i - a_i = \dfrac{b_1 - a_1}{2^{i-1}}$, then we choose a_{k+1} and b_{k+1} as follows. If $f(\dfrac{a_k + b_k}{2}) \geq r$ then set $a_{k+1} = a_k$ and $b_{k+1} = \dfrac{a_k + b_k}{2}$. Otherwise, set $a_{k+1} = \dfrac{a_k + b_k}{2}$ and $b_k = b_{k+1}$, and note that all the aforementioned properties hold when $i = k+1$. Since $\{a_n\}$ is bounded above by b and increasing, we know that $\{a_n\} \to c = \sup(\{a_n\}) \in [a, b]$ by the Monotone Convergence Theorem. Since $\{b_n - a_n\} \to 0$ we know that $\{b_n\} \to c$. Since f is continuous, $\{f(a_n)\} \to f(c)$ and $\{f(b_n)\} \to f(c)$. By the Comparison Theorem, we know $f(c) \leq r$ since $f(a_n) \leq r$ for all n, and $f(c) \geq r$ since $f(b_n) \geq r$ for all n, and therefore $f(c) = r$.

If $f(b) < r < f(a)$ then $-f(a) < -r < -f(b)$ so for some $c \in (a, b)$ we know that $-f(c) = -r$ and therefore $f(c) = r$.

□

Definition 29

Let $f : D \to \mathbb{R}$. We say that f is *uniformly continuous* on $A \subseteq D$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $x, y \in A$ and $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$. We say that f is uniformly continuous if f is uniformly continuous on D.

Theorem 4.18. *Let $f : K \to R$ be continuous, where K is closed and bounded. Then f is uniformly continuous.*

Proof. Suppose f is not uniformly continuous. Then we can pick $\epsilon > 0$ so that for every $\delta > 0$ there are points $x, y \in K$ such that $|x - y| < \delta$ and $|f(x) - f(y)| \geq \epsilon$. For each $n \in \mathbb{N}$ we choose $x_n, y_n \in K$ so that $|x_n - y_n| < \dfrac{1}{n}$ and $|f(x_n) - f(y_n)| \geq \epsilon$. Since $\{x_n\}$ is bounded, by the Bolzano-Weierstrass Theorem there is a subsequence $\{x_{n_k}\} \to p$, where $p \in K$ because K is closed. Since $\{x_{n_k} - y_{n_k}\} \to 0$, we know that $\{y_{n_k}\} \to p$. Since f is continuous, it follows that $\{(f(x_{n_k}))\} \to f(p)$ and $\{f(y_{n_k})\} \to f(p)$. Therefore, $\{f(x_{n_k}) - f(y_{n_k})\} \to 0$. This is impossible since $(-\epsilon, \epsilon)$ excludes $\{f(x_{n_k}) - f(y_{n_k})\}$ for all $k \in \mathbb{N}$. Hence, f is uniformly continuous.

□

Here is an example of a uniformly continuous function.

Example 4.3. *Let $f(x) = 5x$. Prove f is uniformly continuous.*

Solution. Let $\epsilon > 0$. Let $\delta = \dfrac{\epsilon}{5}$. If $|x - y| < \delta$ then $|f(x) - f(y)| = |5x - 5y| = 5|x - y| < 5(\dfrac{\epsilon}{5}) = \epsilon$. Hence, f is uniformly continuous. □

Uniform continuity is a useful property which is stronger than continuity. Continuity at a point p occurs when, for an arbitrary distance $\epsilon > 0$, it is the case that if points x are sufficiently close to p, that is to say with some distance $\delta_p > 0$ of p, their images $f(x)$ are within distance ϵ of $f(p)$. For a uniformly continuous function f, the choice of δ_p can be made the same for every point p in the domain, so there is a uniform choice of δ for a given choice of ϵ (a choice that does not vary with the choice of point p in the domain of f). In the next section we will discuss derivatives, and in one of the exercises we address the fact that if the derivative is bounded for a differentiable function whose domain is an interval then the function is uniformly continuous. This is not an if and only if condition, however. There are functions whose derivatives are unbounded that are still uniformly continuous. The preceding theorem tells us that if we restrict a continuous function to a closed and bounded domain then it is always uniformly continuous, but this does not follow if the domain is not bounded, and it does not follow if the domain is not closed. We leave the case where the domain is bounded but not closed to one of the exercises. Below is an example of a case where a function is not uniformly continuous where the domain is closed but not bounded.

Example 4.4. *Let $f(x) = x^2$. Then prove f is not uniformly continuous on $\mathbb{R}$.*

Solution. Let $\delta > 0$. We know that $(x + \frac{\delta}{2})^2 - x^2 = \delta x + \frac{(\delta)^2}{4} > \delta$. Thus, if $x > \frac{1}{\delta}$ then $|(x + \frac{\delta}{2})^2 - x^2| > \frac{1}{\delta}\delta = 1$, so f is not uniformly continuous. □

Exercises:

Exercise 4.1. *Let $f, g : D \to \mathbb{R}$ be functions so that $\lim\limits_{x\to c} f(x) = M$ and for some $\delta > 0$ it is true that if $|x - c| < \delta$ and $x \in D$ then $|g(x) - L| \leq |f(x) - M|$. Then $\lim\limits_{x\to c} g(x) = L$.*

Exercise 4.2. *Let $f : \mathbb{N} \to \mathbb{R}$ be defined by $f(n) = x_n$ for all $n \in \mathbb{N}$. Then $\lim\limits_{n\to\infty} x_n = L$ if and only if $\{x_n\} \to L$.*

Exercise 4.3. *Let $\lim\limits_{x\to c} f(x) = L$ and let $\epsilon > 0$. If $f(x) = g(x)$ for all $x \in dom(g) \cap (c - \epsilon, c + \epsilon) \setminus \{c\}$ and c is a limit point of the domain of g then $\lim\limits_{x\to c} g(x) = L$.*

Exercise 4.4. *Let $f : D \to \mathbb{R}$ and let c be a limit point of D. Then $\lim\limits_{x\to c} f(x) = L$ if and only if $\lim\limits_{x\to c} f(x) - L = 0$.*

Exercise 4.5. *Let $E \subseteq \mathbb{R}$. Then $\overline{E}$ is closed.*

Exercise 4.6. *Prove Theorem 4.2.*

Exercise 4.7. *Let $f : [a, b] \to [a, b]$ be continuous. Then there is a point $c \in [a, b]$ so that $f(c) = c$.*

Exercise 4.8. *Every polynomial is a continuous function.*

Exercise 4.9. *Let f be continuous at a point c, where $f(c) > 0$. Show that there is a positive number $\delta > 0$ and a positive number $M > 0$ so that $f(x) > M$ for all $x \in (c - \delta, c + \delta) \cap dom(f)$.*

Exercise 4.10. *Give an example, with proof, of a function with bounded domain which is continuous but not uniformly continuous.*

Exercise 4.11. *Give an example, with proof, of a function with bounded range which is continuous but not uniformly continuous. For this example, you may assume that standard trigonometric functions and exponential and logarithmic functions are continuous on their domains (this will be shown in the next chapter).*

Exercise 4.12. *Let $a > 0$ then there is a unique number $c > 0$, the principal nth root of a, having the property that $c^n = a$. We denote this number $c = a^{\frac{1}{n}}$.*

Exercise 4.13. *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then $f([a, b])$ is a closed interval.*

Exercise 4.14. *Let $\{x_n\}$ be a Cauchy sequence in the domain of a uniformly continuous function f. Then $\{f(x_n)\}$ is a Cauchy sequence.*

Exercise 4.15. *Let $f : A \to \mathbb{R}$ be uniformly continuous, where A is bounded. Then $f(A)$ is bounded.*

Exercise 4.16. *Let $f, g : \mathbb{D} \to \mathbb{R}$ be functions with c a limit point of D, so that g is bounded on $(c - \epsilon, c + \epsilon)$ for some $\epsilon > 0$ and $\lim_{x \to c} f(x) = 0$. Then $\lim_{x \to c} f(x)g(x) = 0$.*

Exercise 4.17. *Let f be continuous on (a, b). Then f is uniformly continuous if and only if there is a continuous function $g : [a, b] \to \mathbb{R}$ such that $f(x) = g(x)$ for all $x \in (a, b)$.*

Exercise 4.18. *If $\lim_{x \to c} f(x) = \infty$ then $\lim_{x \to c} \dfrac{1}{f(x)} = 0$.*

Hints:

Hint to Exercise 4.1. *Let $f, g : D \to \mathbb{R}$ be functions so that $\lim_{x \to c} f(x) = M$ and for some $\delta > 0$ it is true that if $|x - c| < \delta$ and $x \in D$ then $|g(x) - L| \leq |f(x) - M|$. Then $\lim_{x \to c} g(x) = L$.*

Try to use the definition of limit directly, picking a distance from c which is less than δ.

Hint to Exercise 4.2. *Let $f : \mathbb{N} \to \mathbb{R}$ be defined by $f(n) = x_n$ for all $n \in \mathbb{N}$. Then $\lim_{n \to \infty} x_n = L$ if and only if $\{x_n\} \to L$.*

Write the definitions of convergence and limit at infinity and compare them.

Hint to Exercise 4.3. *Let $\lim_{x \to c} f(x) = L$ and let $\epsilon > 0$. If $f(x) = g(x)$ for all $x \in dom(g) \cap (c - \epsilon, c + \epsilon) \setminus \{c\}$ and c is a limit point of the domain of g then $\lim_{x \to c} g(x) = L$.*

Write down what the definition of limit being equal to L is for both functions and compare the definitions.

Hint to Exercise 4.4. *Let $f : D \to \mathbb{R}$ and let c be a limit point of D. Then $\lim_{x \to c} f(x) = L$ if and only if $\lim_{x \to c} f(x) - L = 0$.*

Write the definition of each and compare them. Alternately, use the Sequential Characterization of Limits and theorem 3.3.

Hint to Exercise 4.5. *Let $E \subseteq \mathbb{R}$. Then $\overline{E}$ is closed.*

Show that a limit point of $\overline{E}$ is also a limit point of E.

Hint to Exercise 4.6. *Prove Theorem 4.2.*

Parallel the Sequential Characterization of Limits proof.

Hint to Exercise 4.7. *Let $f : [a, b] \to [a, b]$ be continuous. Then there is a point $c \in [a, b]$ so that $f(c) = c$.*

Use the Intermediate Value Theorem on $h(x) = f(x) - x$.

Hint to Exercise 4.8. *Every polynomial is a continuous function.*

Use induction and the fact that the product and sum of continuous functions is continuous.

Hint to Exercise 4.9. *Let f be continuous at a point c, where $f(c) > 0$. Show that there is a positive number $\delta > 0$ and a positive number $M > 0$ so that $f(x) > M$ for all $x \in (c-\delta, c+\delta) \cap dom(f)$.*

Use the definition of continuity to show that for some $\delta > 0$, for all $x \in (c-\delta, c+\delta) \cap dom(f)$ it is true that $|f(x) - f(c)| < \dfrac{|f(c)|}{2}$.

Hint to Exercise 4.10. *Give an example, with proof, of a function with bounded domain which is continuous but not uniformly continuous.*

Consider functions whose tangent line slopes approach infinity or negative infinity.

Hint to Exercise 4.11. *Give an example, with proof, of a function with bounded range which is continuous but not uniformly continuous. For this example, you may assume that standard trigonometric functions and exponential and logarithmic functions are continuous on their domains (this will be shown in the next chapter).*

Consider functions that oscillate a lot.

Hint to Exercise 4.12. *Let $a > 0$ then there is a unique number $c > 0$, the principal nth root of a, having the property that $c^n = a$. We denote this number $c = a^{\frac{1}{n}}$.*

Use the Intermediate Value Theorem.

Hint to Exercise 4.13. *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then $f([a,b])$ is a closed interval.*

Use the Extreme Value Theorem and the Intermediate Value Theorem.

Hint to Exercise 4.14. *Let $\{x_n\}$ be a Cauchy sequence in the domain of a uniformly continuous function f. Then $\{f(x_n)\}$ is a Cauchy sequence.*

Use the definitions of uniform continuity and Cauchy sequence.

Hint to Exercise 4.15. *Let $f : A \to \mathbb{R}$ be uniformly continuous, where A is bounded. Then $f(A)$ is bounded.*

Either use the fact that the uniformly continuous image of a Cauchy sequence is a Cauchy sequence or use a finite collection of points that are spaced so that every point in A is near one of them and the definition of uniform continuity.

Hint to Exercise 4.16. *Let $f, g : \mathbb{D} \to \mathbb{R}$ be functions with c a limit point of D, so that g is bounded on $(c-\epsilon, c+\epsilon)$ for some $\epsilon > 0$ and $\lim\limits_{x\to c} f(x) = 0$. Then $\lim\limits_{x\to c} f(x)g(x) = 0$.*

You can use the definitions directly, or use the Sequential Characterization of Limits and Theorem 3.4.

Hint to Exercise 4.17. *Let f be continuous on (a, b). Then f is uniformly continuous if and only if there is a continuous function $g : [a, b] \to \mathbb{R}$ such that $f(x) = g(x)$ for all $x \in (a, b)$.*

If you assume that f is uniformly continuous on (a, b) then take a sequence $\{x_n\}$ converging to b. Explain why $\{f(x_n)\}$ is a Cauchy sequence and define $g(b)$ to be the point to which this sequence converges, and then prove that g is continuous at b.

Hint to Exercise 4.18. *If $\lim\limits_{x\to c} f(x) = \infty$ then $\lim\limits_{x\to c} \dfrac{1}{f(x)} = 0$.*

Use the definitions, and the fact that the reciprocal of a sufficiently large number can be made as small as we wish.

Solutions:

Solution to Exercise 4.1. *Let $f, g : D \to \mathbb{R}$ be functions so that $\lim_{x\to c} f(x) = M$ and for some $\delta > 0$ it is true that if $|x - c| < \delta$ and $x \in D$ then $|g(x) - L| \leq |f(x) - M|$. Then $\lim_{x\to c} g(x) = L$.*

Proof. Let $\epsilon > 0$. Choose $0 < \gamma < \delta$ so that if $|x - c| < \gamma$ and $x \in D$ then $|f(x) - M| < \epsilon$. Then since $|x - c| < \delta$ it also follows that $|g(x) - L| \leq |f(x) - M| < \epsilon$. Hence, $\lim_{x\to c} g(x) = L$. □

Solution to Exercise 4.2. *Let $f : \mathbb{N} \to \mathbb{R}$ be defined by $f(n) = x_n$ for all $n \in \mathbb{N}$. Then $\lim_{n\to\infty} x_n = L$ if and only if $\{x_n\} \to L$.*

Proof. Assume $\lim_{n\to\infty} x_n = L$. Let $\epsilon > 0$. Then there is some B so that if $n > B$ then $|x_n - L| < \epsilon$. Choose $k \in \mathbb{N}$ so that $k > B$. Then if $n \geq k$ it follows that $|x_n - M| < \epsilon$ so $\{x_n\} \to L$.

Assume $\{x_n\} \to L$. Let $\epsilon > 0$. We can choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - L| < \epsilon$, which means $\lim_{n\to\infty} x_n = L$. □

Solution to Exercise 4.3. *Let $\lim_{x\to c} f(x) = L$ and let $\epsilon > 0$. If $f(x) = g(x)$ for all $x \in dom(g) \cap (c - \epsilon, c + \epsilon) \setminus \{c\}$ and c is a limit point of the domain of g then $\lim_{x\to c} g(x) = L$.*

Proof. Let $\epsilon_1 > 0$. Choose $\delta > 0$ so that if $0 < |x - c| < \delta$ then $|f(x) - L| < \epsilon_1$ if $x \in dom(f)$. Let $\delta_1 = \min\{\epsilon, \delta\}$. Then if $0 < |x - c| < \delta_1$ and $x \in dom(g)$ then $g(x) = f(x)$ so $|g(x) - L| < \epsilon_1$. □

Solution to Exercise 4.4. *Let $f : D \to \mathbb{R}$ and let c be a limit point of D. Then $\lim_{x\to c} f(x) = L$ if and only if $\lim_{x\to c} f(x) - L = 0$.*

Proof. We say that $\lim_{x\to x} f(x) = L$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - L| < \epsilon$.

We say $\lim_{x\to x} f(x) - L = 0$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |x - c| < \delta$ and $x \in D$ then $|f(x) - L - 0| < \epsilon$.

Since these two statements are equivalent, the theorem follows. □

Solution to Exercise 4.5. *Let $E \subseteq \mathbb{R}$. Then $\overline{E}$ is closed.*

Proof. Let p be a limit point of $\overline{E}$ and let (a,b) be an open interval containing p. Then $(a,b)\cap\overline{E}$ contains a point $q\neq p$. If $q\in(\overline{E}\setminus E)$ then q is a limit point of E, which means that (a,b) contains infinitely many points of E, so we can find a point $z\in(a,b)\cap(E\setminus\{p\})$. Since every open interval containing p contains a point of E distinct from p, we know that p is a limit point of E, so $p\in\overline{E}$. Hence, $\overline{E}$ contains all of its limit points and is therefore closed. □

Solution to Exercise 4.6. *Prove Theorem 4.2.*

Proof. First, assume that f is continuous at c. Let $\{x_n\}\subseteq D$ such that $\{x_n\}\to c$. Let $\epsilon>0$. Then for some $\delta>0$, we know that if $|x-c|<\delta$ and $x\in D$ then $|f(x)-f(c)|<\epsilon$. Since $\{x_n\}\to c$, we can find $N\in\mathbb{N}$ so that if $n\geq N$ then $|x_n-c|<\delta$, so if $n\geq N$ then $|x_n-c|<\delta$ which implies that $|f(x_n)-f(c)|<\epsilon$, so $\{f(x_n)\}\to f(c)$.

Next, assume that for every sequence $\{x_n\}\subseteq D$, if $\{x_n\}\to c$ then $\{f(x_n)\}\to f(c)$. Suppose that f is not continuous at c. Then we can find an $\epsilon>0$ so that for every $\delta>0$ there is some $x\in D$ so that $|x-c|<\delta$ but $|f(x)-f(c)|\geq\epsilon$. For each $n\in\mathbb{N}$ we choose $x_n\in D$ so that $|x_n-c|<\dfrac{1}{n}$ and $|f(x_n)-f(c)|\geq\epsilon$. Since $c-\dfrac{1}{n}\leq x_n\leq c+\dfrac{1}{n}$, we know by the Squeeze Theorem that $\{x_n\}\to c$. But $\{f(x_n)\}\not\to f(c)$, contradicting our assumption. Thus, f is continuous at c. □

Solution to Exercise 4.7. *Let $f:[a,b]\to[a,b]$ be continuous. Then there is a point $c\in[a,b]$ so that $f(c)=c$.*

Proof. If $f(a)=a$ or $f(b)=b$ then we have a fixed point as desired. Assume $f(a)>a$ and $f(b)<b$. Let $h(x)=f(x)-x$. Since x is continuous and f is continuous, we know that h is continuous. Also, $h(a)>0>h(b)$, so by the Intermediate Value Theorem there is some $c\in(a,b)$ so that $h(c)=0$ which means that $f(c)=c$. □

Solution to Exercise 4.8. *Every polynomial is a continuous function.*

Proof. We will proceed by induction on the degree of the polynomial. Let $\epsilon>0$. We know a degree one polynomial $f(x)=ax+b$ is continuous by Theorem 4.3 and Theorem 4.7.

Assume that all polynomials of degree k are continuous. Let $P(x)=a_{k+1}x^{k+1}+a_kx^k+\ldots+a_1x+a_0$ be a polynomial of degree $k+1$. Then we know that $a_{k+1}x^k$, x, and $a_kx^k+\ldots+a_1x+a_0$ are each continuous from the induction hypothesis, so by theorem 4.7 we conclude that $P(x)$ is continuous since $P(x)=(a_{k+1}x^k)(x)+(a_kx^k+\ldots+a_1x+a_0)$. By induction, the result follows. □

Solution to Exercise 4.9. *Let f be continuous at a point c, where $f(c)>0$. Show that there is a positive number $\delta>0$ and a positive number $M>0$ so that $f(x)>M$ for all $x\in(c-\delta,c+\delta)\cap dom(f)$.*

Proof. Choose $\delta > 0$ so that if $|x-c| < \delta$ and $x \in dom(f)$ then $|f(x)-f(c)| < \dfrac{f(c)}{2}$. Then by Theorem 1.17, we know that $f(x) > f(c) - \dfrac{f(c)}{2} = \dfrac{f(c)}{2}$ for all $x \in (c-\delta, c+\delta) \cap dom(f)$. □

Solution to Exercise 4.10. *Give an example, with proof, of a function with bounded domain which is continuous but not uniformly continuous.*

Proof. Let $f(x) = \dfrac{1}{x}$ on $(0,1)$. Then f is continuous by Theorem 4.7, but if we set $\epsilon = 1$ and choose any $\delta > 0$ we can find $\dfrac{1}{k} < \delta$ and then $|\dfrac{1}{k} - \dfrac{1}{k+1}| < \delta$, but $|f(\dfrac{1}{k}) - f(\dfrac{1}{k+1})| = |k+1-k| = 1$. Thus, f is not uniformly continuous. □

Solution to Exercise 4.11. *Give an example, with proof, of a function with bounded range which is continuous but not uniformly continuous. For this example, you may assume that standard trigonometric functions and exponential and logarithmic functions are continuous on their domains (this will be shown in the next chapter).*

Proof. Let $f(x) = \sin(\dfrac{1}{x})$ on $(0,1)$. We know f is a composition of a continuous function and a ratio of two continuous functions and is continuous by Theorems 4.12 and 4.7. For any $\delta > 0$ we can choose $\dfrac{1}{2\pi n} < \delta$ so that $|\dfrac{1}{2\pi n + \frac{\pi}{2}} - \dfrac{1}{2\pi n + \frac{3\pi}{2}}| < \delta$, but $|f(\dfrac{1}{2\pi n + \frac{\pi}{2}}) - f(\dfrac{1}{2\pi n + \frac{3\pi}{2}})| = |1-(-1)| = 2$. Thus, f is not uniformly continuous. □

Solution to Exercise 4.12. *Let $a > 0$ then there is a unique number $c > 0$, the nth root of a, having the property that $c^n = a$.*

Proof. Let $f(x) = x^n$. Then we know f is continuous by Exercise 4.8. Also, $f(0) = 0 < a$ and $(a+1)^n = 1 + na + ... + a^n > a$ by the Binomial Theorem. Hence, by the Intermediate Value Theorem, for some $c \in (0, a+1)$ it follows that $f(c) = a$, or in other words, $c^n = a$. The fact that this number is unique follows from Exercise 2.13 which shows that $f(x)$ is increasing on $[0,\infty)$ and therefore one to one. □

Solution to Exercise 4.13. *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then $f([a,b])$ is a closed interval.*

Proof. By the Extreme Value Theorem there are points $s, t \in [a,b]$ so that $f(s) \leq f(x) \leq f(t)$ for all $x \in [a,b]$. By the Intermediate Value Theorem, for any $k \in (f(s), f(t))$ there is a point c between s and t so that $f(c) = k$. Hence, the set of points in $f([a,b])$ is exactly the interval $[f(s), f(t)]$. □

Solution to Exercise 4.14. *Let $\{x_n\}$ be a Cauchy sequence in the domain of a uniformly continuous function f. Then $\{f(x_n)\}$ is a Cauchy sequence.*

Proof. Let $\epsilon > 0$. Choose $\delta > 0$ so that if $|x-y| < \delta$ and $x, y \in dom(f)$ then $|f(x)-f(y)| < \epsilon$. Choose $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|x_n - x_m| < \delta$. Then if $n, m \geq k$ we know that $|f(x_n) - f(x_m)| < \epsilon$, so $\{f(x_n)\}$ is a Cauchy sequence. □

Solution to Exercise 4.15. *Let $f : A \to \mathbb{R}$ be uniformly continuous, where A is bounded. Then $f(A)$ is bounded.*

Proof. Suppose f is not bounded. Then for each $n \in \mathbb{N}$ we can choose $x \in A$ so that $|f(x)| > n$. Since A is bounded we know $\{x_n\}$ is bounded and has a convergent subsequence $\{x_{n_i}\}$, so $\{x_{n_i}\}$ is a Cauchy sequence. By Exercise 4.14, this means that $\{f(x_{n_i})\}$ is a Cauchy sequence, but this is impossible since $\{f(x_{n_i})\}$ is not bounded because for any $M > 0$ we can find $k \in \mathbb{N}$ so that $k > M$, so $|f(x_{n_k})| > n_k \geq k > M$ by Theorem 3.12. Hence, $f(A)$ is bounded. □

Solution to Exercise 4.16. *Let $f, g : \mathbb{D} \to \mathbb{R}$ be functions with c a limit point of D, so that g is bounded on $(c-\epsilon, c+\epsilon)$ for some $\epsilon > 0$ and $\lim_{x\to c} f(x) = 0$. Then $\lim_{x\to c} f(x)g(x) = 0$.*

Proof. Let $\{x_n\} \to c$ where $\{x_n\} \subseteq D \setminus \{c\}$. Then by the Sequential Characterization of Limits we know that $\{f(x_n)\} \to 0$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $x_n \in (c-\epsilon, c+\epsilon)$. Then by Theorem 3.4, we know that $\{f(x_{n+k})g(x_{n+k})\} \to 0$, so by Theorem 3.26 we know that $\{f(x_n)g(x_n)\} \to 0$. Thus, by the Sequential Characterization of Limits again, we know that $\lim_{x\to c} f(x)g(x) = 0$.

□

Alternate proof:

Proof. Since g is bounded we can find $M > 0$ so that $|g(x)| \leq M$ for all $x \in D \cap (c-\epsilon, c+\epsilon)$. Let $\epsilon_1 > 0$. Choose $0 < \delta < \epsilon$ so that if $0 < |x-c| < \delta$ and $x \in D$ then $|f(x)| < \frac{\epsilon_1}{M}$. If $0 < |x-c| < \delta$ and $x \in D$ then $|g(x)f(x)| \leq \frac{\epsilon_1}{M}M = \epsilon_1$, which means that $\lim_{x\to c} f(x)g(x) = 0$. □

Solution to Exercise 4.17. *Let f be continuous on (a, b). Then f is uniformly continuous if and only if there is a continuous function $g : [a, b] \to \mathbb{R}$ such that $f(x) = g(x)$ for all $x \in (a, b)$.*

Proof. First, assume that g exists. Then since g is continuous on $[a, b]$ we know that g is uniformly continuous, which implies that f is uniformly continuous.

Next, assume that f is uniformly continuous. Let $\{x_n\} \to a$, where $\{x_n\} \subset (a, b)$. Then $\{x_n\}$ is a Cauchy sequence, so $\{f(x_n)\}$ is a Cauchy sequence by Exercise 4.14, which means that $\{f(x_n)\}$ converges to a point which we will define to be $g(a)$ (by Theorem 3.25).

Let $\{y_n\} \subset (a, b)$ so that $\{y_n\} \to a$. Let $\epsilon > 0$. Since f is uniformly continuous, we can choose $\delta > 0$ so that if $|x-y| < \delta$ then $|f(x) - f(y)| < \epsilon$. Since $\{x_n\} \to a$ and $\{y_n\} \to a$

we know that $\{x_n - y_n\} \to 0$. Thus, we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - y_n| < \delta$, so $|f(x_n) - f(y_n)| < \epsilon$. Hence, $\{f(x_n) - f(y_n)\} \to 0$, which means that $\{f(y_n)\} \to g(a)$ by Exercise 4.10. This implies that $\lim\limits_{x \to a} f(x) = L$ by the Sequential Characterization of Limits, so g is continuous at a. We similarly assign $g(b)$ by using the same process (replace a by b in the preceding argument). Hence, $g(x) = f(x)$ for $x \in (a,b)$, with $g(a)$ and $g(b)$ thus defined, is continuous.

□

Solution to Exercise 4.18. *If* $\lim\limits_{x \to c} f(x) = \infty$ *then* $\lim\limits_{x \to c} \dfrac{1}{f(x)} = 0$.

Proof. Let $\epsilon > 0$. Since $\lim\limits_{x \to c} f(x) = \infty$, we can choose $\delta > 0$ so that if $0 < |x - c| < \delta$ then $f(x) > \dfrac{1}{\epsilon}$, which means that $0 < \dfrac{1}{f(x)} < \epsilon$ and thus $\lim\limits_{x \to c} \dfrac{1}{f(x)} = 0$. □

Chapter 5

Differentiation

Definition 30

Let $f : D \to \mathbb{R}$ and let $x_0 \in D^\circ$. We say that $y = f(x)$ is *differentiable* at x_0 if $\lim_{h\to 0} \dfrac{f(x_0+h) - f(x_0)}{h}$ exists, in which case we call this limit the *derivative* of f at x_0, denoted $f'(x_0)$ or $\dfrac{dy}{dx}(x_0)$.

We let $f''(x)$ denote $(f'(x))'$, and let $f'''(x) = (f''(x))'$ and so on. We also use the notation $f^{(i)}(x)$ to denote the ith derivative of f at x (in other words, we define $f^{(i)}(x)$ recursively by stating that $f^{(i)}(x) = (f^{(i-1)}(x))'$ for natural numbers $i > 1$, where $f^{(1)}(x) = f'(x)$).

Note that if a function is defined on a point in the interior of its domain then the definition of limit and continuity of a function can omit the "and $x \in D$" portion of the definition because by choosing δ small enough, it is guaranteed that all x so that $|x - c| < \delta$ will be in the domain of the function. For a function which is differentiable at a point p, the point p is always in the interior of the domain.

Theorem 5.1. *For any set D it is true that $x_0 \in D^\circ$ if and only if $x_0 \in (a, b) \subseteq D$ for some open interval (a, b).*

Proof. We know that $x_0 \in D^\circ$ if and only if x_0 is contained in the open interval $(x_0 - \epsilon, x_0 + \epsilon) \subseteq D$ for some $\epsilon > 0$. If $x_0 \in (a, b) \subseteq D$ for some open interval (a, b) then since (a, b) is an open set by Theorem 3.2, we can find $\epsilon > 0$ so that $x_0 \in (x_0 - \epsilon, x_0 + \epsilon) \subseteq (a, b) \subseteq D$ which means that $x \in D^\circ$. □

Theorem 5.2. *Let $f : (a, b) \to \mathbb{R}$ and let $c \in (a, b)$. Then $f'(c)$ exists if and only if $f'(c) = \lim_{x\to c} \dfrac{f(x) - f(c)}{x - c}$.*

Proof. We know $f'(c) = \lim_{h\to 0} \dfrac{f(c+h) - f(c)}{h}$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |h - 0| < \delta$ then $|\dfrac{f(c+h) - f(c)}{h} - f'(c)| < \epsilon$ which is true if and only if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |x - c| < \delta$ then $|\dfrac{f(c + x - c) - f(c)}{x - c} - f'(c)| < \epsilon$, which is true if and only if $f'(c) = \lim_{x \to c} \dfrac{f(x) - f(c)}{x - c}$.

□

Theorem 5.3. *Let $f : (a, b) \to \mathbb{R}$ and let $c \in (a, b)$. If f is differentiable at c then f is continuous at c.*

Proof. Since $\lim_{x \to c} f(x) - f(c) = \lim_{x \to c} \dfrac{f(x) - f(c)}{x - c}(x - c) = f'(c)(0) = 0$, it follows that $\lim_{x \to c} f(x) = f(c)$, so f is continuous at c.

□

There are other ways of looking at differentiability that are sometimes instructive.

Theorem 5.4. *Let $f(x)$ be defined on an open interval containing x_0. Then f is differentiable at x_0 if and only if there is a function $F : dom(f) \to \mathbb{R}$ so that F is continuous at x_0 and $f(x) = F(x)(x - x_0) + f(x_0)$ for all $x \in dom(f)$, in which case $F(x_0) = f'(x_0)$.*

Proof. First, assume that $f'(x_0)$ exists. Define $F(x) = \dfrac{f(x) - f(x_0)}{x - x_0}$ if $x \neq x_0$ and define $F(x_0) = f'(x_0)$. Since $\lim_{x \to x_0} F(x) = \lim_{x \to x_0} \dfrac{f(x) - f(x_0)}{x - x_0} = f'(x_0) = F(x_0)$ we know that F is continuous at x_0. For $x \in dom(f) \setminus \{x_0\}$ we have that $F(x)(x - x_0) + f(x_0) = \dfrac{f(x) - f(x_0)}{x - x_0}(x - x_0) + f(x_0) = f(x)$. For $x = x_0$, we see $F(x)(x - x_0) + f(x_0) = f'(x_0)(0) + f(x_0) = f(x)$ as well. Thus, $f(x) = F(x)(x - x_0) + f(x_0)$ for all $x \in dom(f)$.

Next, assume that there is a function $F : dom(f) \to \mathbb{R}$ so that F is continuous at x_0 and $f(x) = F(x)(x - x_0) + f(x_0)$ for all $x \in dom(f)$. Solving for F, we get that $F(x) = \dfrac{f(x) - f(x_0)}{x - x_0}$ if $x \neq x_0$. Since F is continuous at x_0 we know that $\lim_{x \to x_0} F(x) = F(x_0)$ which means that $\lim_{x \to x_0} \dfrac{f(x) - f(x_0)}{x - x_0} = F(x_0)$, so by definition $F(x_0) = f'(x_0)$ and f is differentiable at x_0. □

Theorem 5.5. *Let $f(x)$ be defined on an open interval containing x_0. Then f is differentiable at x_0 if and only if there is a function $\epsilon(\Delta x)$ defined for all Δx so that $x_0 + \Delta x \in dom(f)$, and a constant k, so that $\lim_{\Delta x \to 0} \dfrac{\epsilon(\Delta x)}{\Delta x} = 0$ and $f(x_0 + \Delta x) - f(x_0) = k\Delta x + \epsilon(\Delta x)$ for all $x \in dom(f)$, in which case $k = f'(x_0)$.*

Proof. First, assume that f is differentiable at x_0. Then set $\epsilon(\Delta x) = f(x_0 + \Delta x) - f(x_0) - f'(x_0)\Delta x$ we have that $\lim_{\Delta x \to 0} \dfrac{\epsilon(\Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \dfrac{f(x_0 + \Delta x) - f(x_0) - f'(x_0)\Delta x}{\Delta x} = \lim_{\Delta x \to 0} \dfrac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} - f'(x_0) = f'(x_0) - f'(x_0) = 0$.

Next, assume that there is a function $\epsilon(\Delta x)$ defined for all Δx so that $x_0 + \Delta x \in dom(f)$, and a constant k, so that $\lim\limits_{\Delta x \to 0} \dfrac{\epsilon(\Delta x)}{\Delta x} = 0$ and $f(x_0 + \Delta x) - f(x_0) = k(x - x_0) + \epsilon(\Delta x)$ for all $x \in dom(f)$. Then, solving for $\epsilon(\Delta x) = f(x_0 + \Delta x) - f(x_0) - f'(x_0)\Delta x$, we know that $\lim\limits_{\Delta x \to 0} \dfrac{\epsilon(\Delta x)}{\Delta x} = 0$, so $= \lim\limits_{\Delta x \to 0} \dfrac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} - f'(x_0) = 0$, so $\lim\limits_{\Delta x \to 0} \dfrac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = f'(x_0)$. □

Sometimes it is also helpful to rephrase the preceding theorem as follows, depending on the context.

Theorem 5.6. *Let $f(x)$ be defined on an open interval containing x_0. Then f is differentiable at x_0 if and only if there is a function $\delta(\Delta x)$ defined for all Δx so that $x_0 + \Delta x \in dom(f)$, and a constant k, so that $\lim\limits_{\Delta x \to 0} \delta(\Delta x) = 0$ and $f(x_0 + \Delta x) - f(x_0) - k\Delta x = \delta(\Delta x)\Delta x$ for all $x \in dom(f)$ (or, in other words, $\dfrac{f(x_0 + \Delta x) - f(x_0) - k\Delta x}{\Delta x} = \delta(\Delta x)$), in which case $k = f'(x_0)$.*

Proof. Set $\delta(\Delta x) = \dfrac{\epsilon(\Delta x)}{\Delta x}$ in Theorem 5.5 and the result follows. □

The following is another way to look at differentiation in a more geometric context.

Theorem 5.7. *Let $f(x)$ be defined on an open interval containing x_0. Then f is differentiable at x_0 if and only if there is a number m so that for any numbers $c < m < d$, there is a $\delta > 0$ so that if $|x - x_0| < \delta$ then $f(x)$ is between or equal to $c(x) = c(x - x_0) + f(x_0)$ and $d(x) = d(x - x_0) + f(x_0)$, in which case $m = f'(x_0)$.*

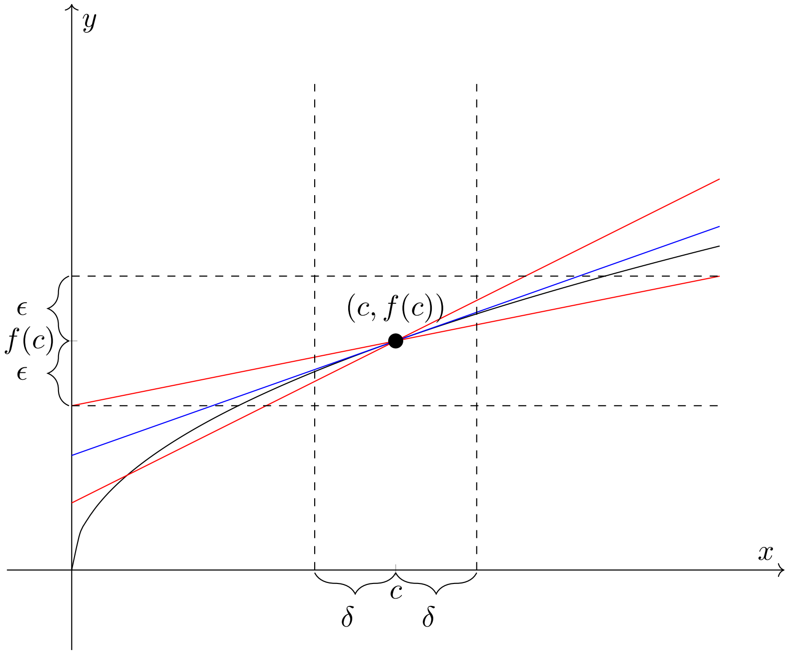
Proof. Assume $f'(x_0)$ exists and set $m = f'(x_0)$. Let $\epsilon = \min\{m - c, d - m\}$. Choose $\delta > 0$ so that if $|x - x_0| < \delta$ then $|\dfrac{f(x) - f(x_0)}{x - x_0} - m| < \epsilon$, so $c < \dfrac{f(x) - f(x_0)}{x - x_0} < d$. If $x > x_0$ then $c(x - x_0) < f(x) - f(x_0) < d(x - x_0)$, which means that $c(x - x_0) + f(x_0) < f(x) < d(x - x_0) + f(x_0)$, and if $x < x_0$ then $c(x - x_0) + f(x_0) > f(x) > d(x - x_0) + f(x_0)$.

Assume there is a number m so that for any numbers $c < m < d$, there is a $\delta > 0$ so that if $|x - x_0| < \delta$ then $f(x)$ is between or equal to $c(x) = c(x - x_0) + f(x_0)$ and $d(x) = d(x - x_0) + f(x_0)$. Let $\epsilon > 0$. Choose $c = m - \epsilon$ and $d = m + \epsilon$. Then there is a $\delta > 0$ so that if $|x - x_0| < \delta$ then $f(x)$ is between or equal to $c(x) = c(x - x_0) + f(x_0)$ and $d(x) = d(x - x_0) + f(x_0)$. If $x > x_0$ then $c(x - x_0) < d(x - x_0)$, so $c(x - x_0) + f(x_0) < f(x) < d(x - x_0) + f(x_0)$, and $m - \epsilon < \dfrac{f(x) - f(x_0)}{x - x_0} < m + \epsilon$, so $-\epsilon < \dfrac{f(x) - f(x_0)}{x - x_0} - m < \epsilon$, which means that $|\dfrac{f(x) - f(x_0)}{x - x_0} - m| < \epsilon$. Similarly, if $x < x_0$ then $c(x - x_0) > d(x - x_0)$, so $c(x - x_0) + f(x_0) > f(x) > d(x - x_0) + f(x_0)$, which means $-\epsilon < \dfrac{f(x) - f(x_0)}{x - x_0} - m < \epsilon$, so $|\dfrac{f(x) - f(x_0)}{x - x_0} - m| < \epsilon$. Hence, $\lim\limits_{x \to x_0} \dfrac{f(x) - f(x_0)}{x - x_0} = m = f'(x_0)$.

□

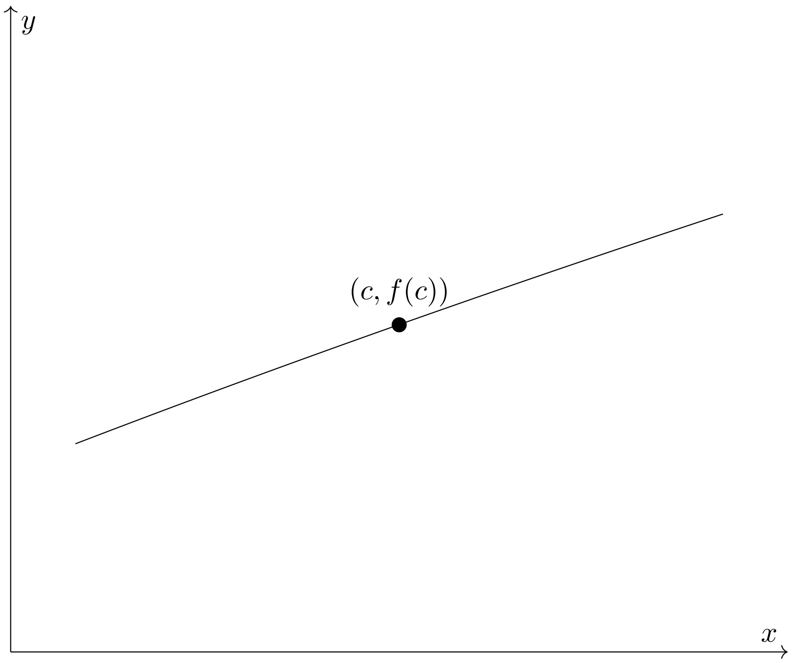
The last of these forms a derivative helps us to picture a derivative visually. A function is differentiable at a point $(c, f(c))$ if there is a tangent line with slope $f'(c)$ through that point so that if you take a line of greater slope though $(c, f(c))$ and a line of smaller slope through $(c, f(c))$ then there is a $\delta > 0$ distance about c so that all the points of the graph of $y = f(x)$ that have x values between $c - \delta$ and $c + \delta$ have corresponding y values between the two lines given. Sometimes this is thought of in terms of angles. If you take any double cone of any positive angle between the lines forming the cone with vertex at $(c, f(c))$ where the tangent line bisects the angle, then a vertical band about $x = c$ which is sufficiently narrow will have all points in the graph within that band sandwiched inside the double cone. In other words, a differentiable function is approximately flat near a point on the curve with x-value at a point where the function is differentiable. If you were to zoom in on a portion of the curve of sufficiently small diameter then the magnified image would look more and more like the tangent line.

Differentiability at c

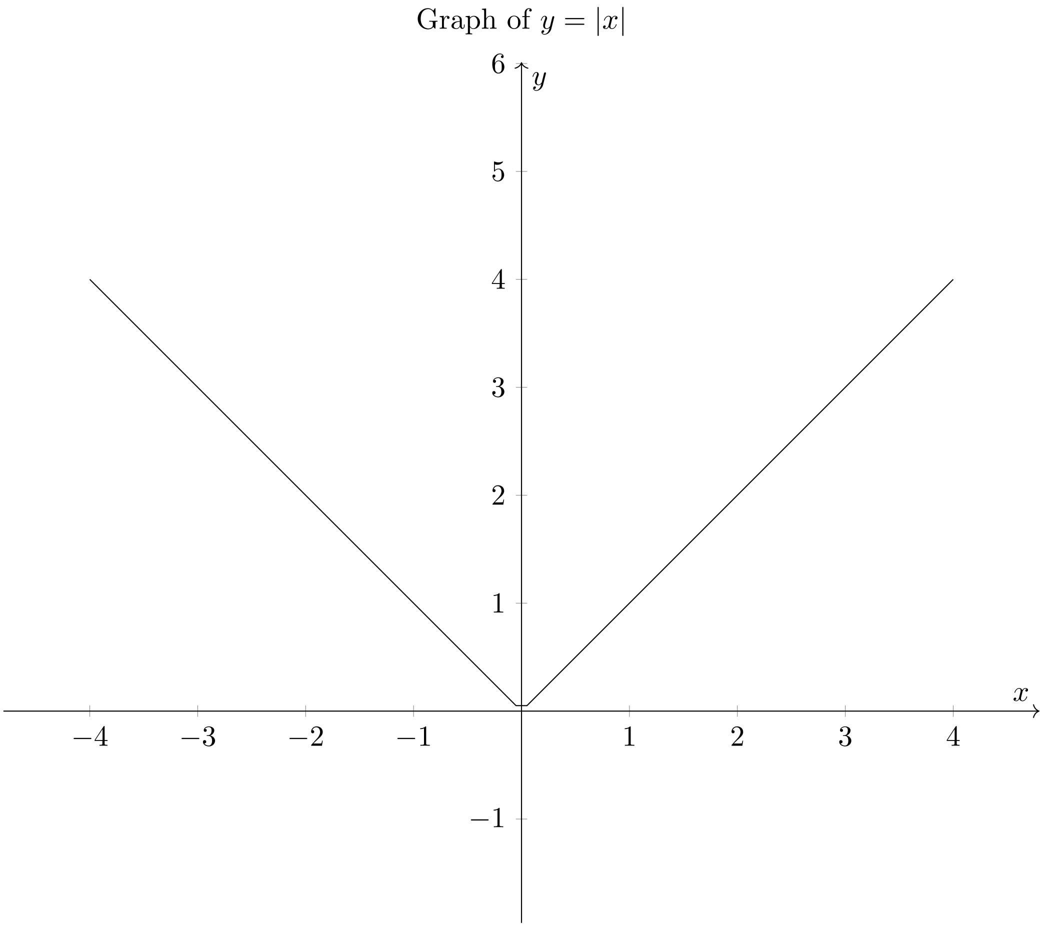


Magnifying the region near $(c, f(c))$ in the picture above, the graph looks like this.

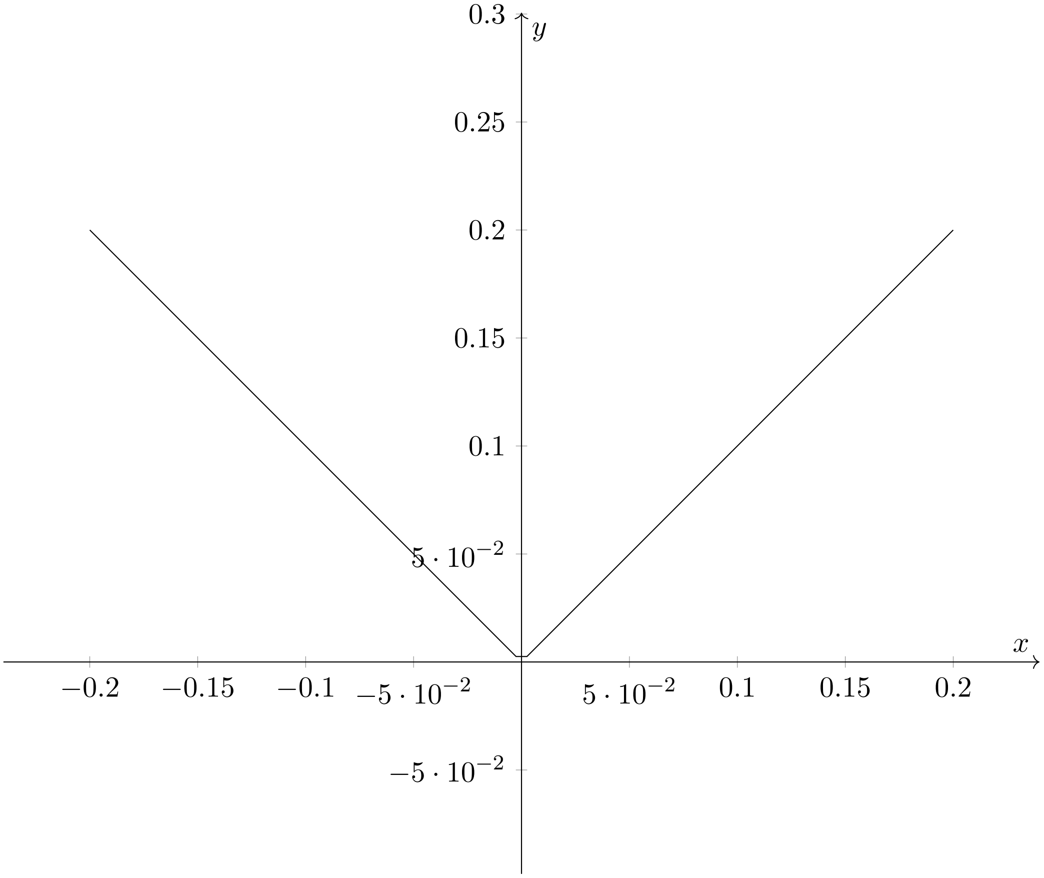
Magnified Graph near c



In contrast, $f(x) = |x|$ is not differentiable at the origin. If you were to zoom in on the vertex of the graph it would always look like a V shape, no matter how much you magnified the image. The graph would never look flat and linear because the function is not differentiable.


Graph of $y = |x|$
6
y
5
4
3
2
1
x
-4
-3
-2
-1
1
2
3
4
-1

Magnifying Graph of $y = |x|$ Near Origin



Theorem 5.8. *Let f, g be differentiable at c. Then there is an open interval (a, b) so that $c \in (a, b) \subseteq dom(f) \cap dom(g)$.*

Proof. By the definition of differentiable there $\epsilon_1, \epsilon_2 > 0$ so that $(c - \epsilon_1, c + \epsilon_1) \subseteq dom(f)$ and $(c - \epsilon_2, c + \epsilon_2) \subseteq dom(g)$. Letting $\epsilon = \min(\epsilon_1, \epsilon_2)$ we see that $(c - \epsilon, c + \epsilon) \subseteq dom(f) \cap dom(g)$. □

Theorem 5.9. *Let f, g be differentiable at x. Then:*

(a) $(f + g)'(x) = f'(x) + g'(x)$

(b) $(fg)'(x) = f(x)g'(x) + g(x)f'(x)$

(c) If $g(x) \neq 0$ then $(\frac{f}{g})'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$.

Proof. By Theorem 5.8 we know that x is contained in an open interval contained in the domains of $f + g$ and fg, and by Theorem 4.5, if $g(x) \neq 0$ then $\frac{f}{g}$ is also defined on an open interval containing x.

(a) $(f + g)'(x) = \lim_{h \to 0} \frac{f(x+h) + g(x+h) - (f(x) + g(x))}{h} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} + \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} = f'(x) + g'(x)$ by the sum rule for limits.

(b) $(fg)'(x) = \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} =$
$\lim_{h\to 0} \frac{f(x+h)g(x+h) - g(x+h)f(x) + g(x+h)f(x) - f(x)g(x)}{h}$
$= \lim_{h\to 0} g(x+h)\frac{f(x+h) - f(x)}{h} + \lim_{h\to 0} f(x)\frac{g(x+h) - g(x)}{h} = f(x)g'(x) + g(x)f'(x)$ by and Theorems 4.6, 4.11 and 4.4.

(c) $(\frac{f}{g})'(x) = \lim_{h\to 0} \frac{\frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}}{h} = \lim_{h\to 0} \frac{f(x+h)g(x) - f(x)g(x+h)}{g(x+h)g(x)h}$
$= \lim_{h\to 0} \frac{1}{g(x+h)g(x)} \; \frac{f(x+h)g(x) - f(x)g(x) + f(x)g(x) - f(x)g(x+h)}{h}$
$= \lim_{h\to 0} \frac{1}{g(x+h)g(x)} g(x)\frac{f(x+h) - f(x)}{h} - \lim_{h\to 0} \frac{1}{g(x+h)g(x)} f(x)\frac{g(x+h) - g(x)}{h}$
$= \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$ by Theorems 4.6, 4.11 and 4.4, assuming that $g(x) \neq 0$.

□

Theorem 5.10. *If $f(x) = x$ (on some open interval I) then $f'(x) = 1$ for each $x \in I$. If $f(x) = k$ (one some open interval) then $f'(x) = 0$ for each $x \in I$.*

Proof. $\lim_{h\to 0} \frac{x+h-x}{h} = 1$ and $\lim_{h\to 0} \frac{k-k}{h} = 0$ by Theorem 4.3.

□

Theorem 5.11. *If f is differentiable at x_0 and g is differentiable at $f(x_0)$ then $g \circ f$ is defined on an open interval containing x_0.*

The proof of this theorem is left as an exercise.

For most examples the chain rule is a straightforward idea. If you are traveling at 10 miles per hour currently, and you are painting a road requiring one half of one kilogram of paint per mile then you are using 5 kilograms of paint per hour. In this manner, if you have $p = p(s)$ is paint as a function of displacement and $s = s(t)$ is displacement as a function of time then $\frac{dp}{ds}\frac{ds}{dt} = \frac{dp}{dt}$. The amount paint changes per unit of distance times the amount distance changes per unit of time is the amount that paint changes per unit of time. This is the idea behind the chain rule. Some people think of it as cancelling the ds in the fraction above. This, of course, is nonsense, because the symbol $\frac{ds}{dt}$ refers to a limit of a difference quotient ratio of change in position over change in time, and not an actual fraction. There are ways of interpreting decimals as fractions of differentials and ways of formalizing the operations that correspond to what would have been such symbol cancellations which make sense, but these would have to be developed (and generally follow from the chain rule). Even so, the idea of the chain rule corresponds to the idea of multiplying rates of change as described.

We begin with an inadequate proof of the chain rule which is simple. We include it because it is correct in the vast majority of cases and it is easy to understand. With the restrictions we place on the function, the proof is correct.

Theorem 5.12. *Special Case of Chain Rule. Let f be differentiable at some point x_0 so that for some $\epsilon > 0$ it is true that $f(x) \neq f(x_0)$ for all $x \neq x_0$ so that $x \in (x_0 - \epsilon, x_0 + \epsilon)$. Let g be differentiable at $f(x_0)$. Then $(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$.*

Proof. We set $y = f(x)$ and $y_0 = f(x_0)$. Then for $|h| < \epsilon$ we have $\lim_{x \to x_0} \dfrac{g(f(x)) - g(f(x_0))}{x - x_0} =$
$\lim_{x \to x_0} \dfrac{g(y) - g(y_0)}{y - y_0} \dfrac{y - y_0}{x - x_0} = \lim_{y \to y_0} \dfrac{g(y) - g(y_0)}{y - y_0} \lim_{x \to x_0} \dfrac{f(x) - f(x_0)}{x - x_0} = g'(y_0)f'(x_0) = g'(f(x_0))f'(x_0)$
by Theorem 4.11 since g is continuous at y_0 and $f(x) - f(x_0) \neq 0$. □

This is not the full form of the chain rule. It is appealing, and works for most functions. To prove the chain rule properly, however, we need to deal with the fact that $\dfrac{g(f(x)) - g(f(x_0))}{x - x_0}$ cannot be written as $\dfrac{g(y) - g(y_0)}{y - y_0} \dfrac{y - y_0}{x - x_0}$ when $y = y_0$, which could happen at points arbitrarily close to x_0 potentially. This can be done by using the form of differentiability discussed in Theorem 5.4, which essentially creates a continuous function to fill in the holes in the cases where $y = y_0$.

Theorem 5.13. *Chain Rule. If f is differentiable at a and g is differentiable at $f(a)$ then $(g \circ f)'(a) = g'(f(a))f'(a)$.*

Proof. We know $g \circ f$ is defined on an open interval containing a by Theorem 5.11.

By Theorem 5.4, there is a function $G : dom(g) \to \mathbb{R}$ which is continuous at $f(a)$ so that $g(x) = G(x)(x - f(a)) + g(f(a))$ for all $x \in dom(g)$ and $G(f(a)) = g'(f(a))$.

We can write $(g \circ f)'(a) = \lim_{x \to a} \dfrac{g(f(x)) - g(f(a))}{x - a} =$
$\lim_{x \to a} \dfrac{G(f(x))(f(x) - f(a)) + g(f(a)) - g(f(a))}{x - a} =$
$\lim_{x \to a} G(f(x)) \dfrac{(f(x) - f(a))}{x - a} = G(f(a))f'(a) = g'(f(a)f'(a)$ since G is continuous at $f(a)$.
□

Definition 31

Let f be a real valued function defined at a point c. We say that $(c, f(c))$ is a *local maximum* for f if there is an $\epsilon > 0$ so that $(c - \epsilon, c + \epsilon) \subset dom(f)$ and $f(x) \leq f(c)$ for each $x \in (c - \epsilon, c + \epsilon)$. We say that $(c, f(c))$ is a *local minimum* for f if there is an $\epsilon > 0$ so that for each $x \in (c - \epsilon, c + \epsilon)$ it is true that $f(x) \geq f(c)$. If $(c, f(c))$ is a local maximum or a local minimum then $(c, f(c))$ is a *local extremum*.

Theorem 5.14. *Fermat's Theorem. Let $f : (a, b) \to \mathbb{R}$ be differentiable and let $(t, f(t))$ be a local extremum for f. Then $f'(t) = 0$.*

Proof. First, assume that $(t, f(t))$ is a local maximum. Then there is an $\epsilon > 0$ so that $(t - \epsilon, t + \epsilon) \subset dom(f)$ and $f(t) \geq f(x)$ for all $x \in (t - \epsilon, t + \epsilon)$. Choose sequences $\{x_n\} \subset (t - \epsilon, t)$ and $\{y_n\} \subset (t, t + \epsilon)$ which both converge to t. For each $n \in \mathbb{N}$ we know $\dfrac{f(x_n) - f(t)}{x_n - t} \geq 0$ and $\dfrac{f(y_n) - f(t)}{y_n - t} \leq 0$, so by the Comparison Theorem it follows that $\lim\limits_{x \to t} \dfrac{f(x) - f(t)}{x - t} = \lim\limits_{n \to \infty} \dfrac{f(x_n) - f(t)}{x_n - t} \geq 0$, and $\lim\limits_{x \to t} \dfrac{f(x) - f(t)}{x - t} = \lim\limits_{n \to \infty} \dfrac{f(y_n) - f(t)}{y_n - t} \leq 0$, so $\lim\limits_{x \to t} \dfrac{f(x) - f(t)}{x - t} = 0$ by the Sequential Characterization of Limits.

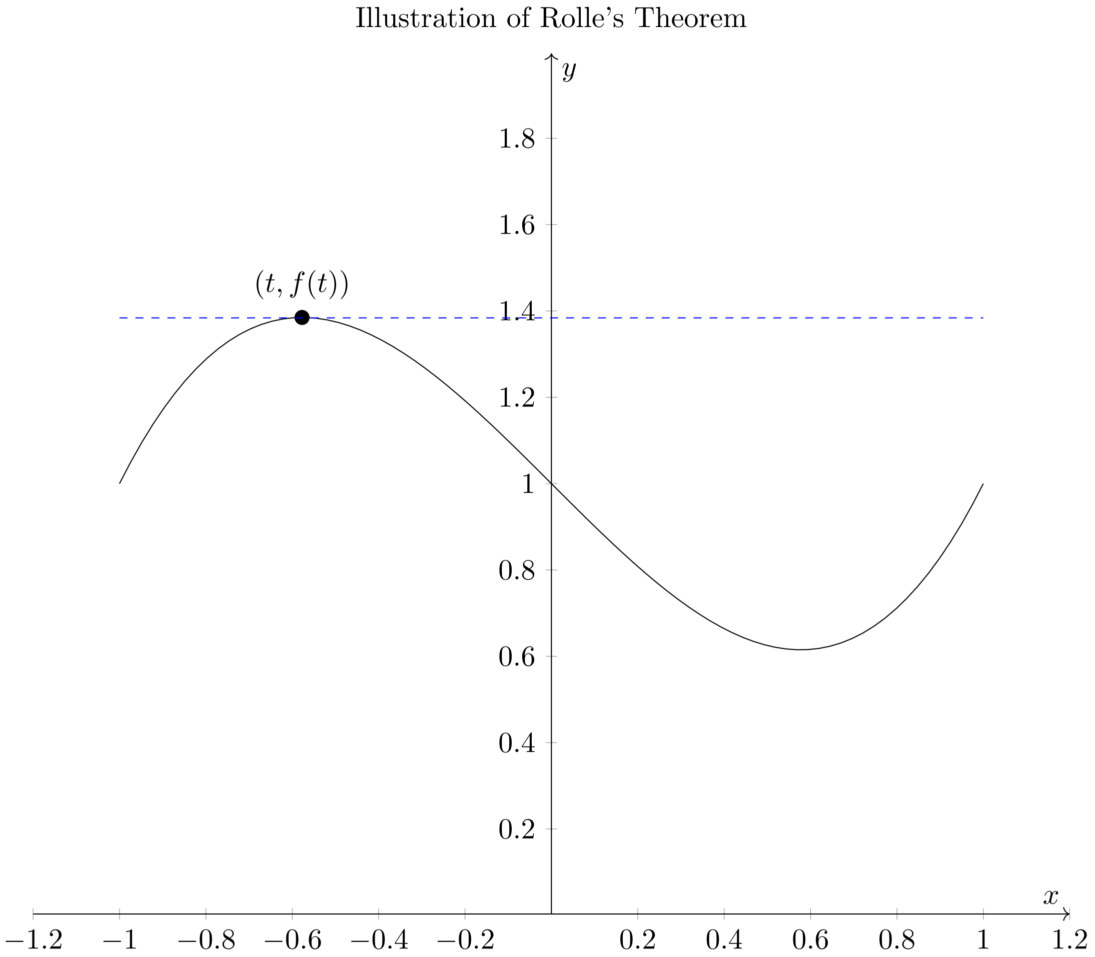
If $(t, f(t))$ is a local minimum then there is an $\epsilon > 0$ so that $(t - \epsilon, t + \epsilon) \subset dom(f)$ and $f(t) \leq f(x)$ so $-f(t) \geq -f(x)$ for all $x \in (t - \epsilon, t + \epsilon)$. Thus, $(t, -f(t))$ is a local maximum for the function $-f$, so $-f'(t) = 0$ and hence $f'(t) = 0$.

□

Theorem 5.15. *Rolle's Theorem. If f is continuous on $[a, b]$ and differentiable on (a, b) and $f(a) = f(b)$ then there is a point $c \in (a, b)$ so that $f'(c) = 0$.*

Proof. By the Extreme Value Theorem we can find $s, t \in [a, b]$ so that $f(s) \leq f(x) \leq f(t)$ for all $x \in [a, b]$. If $f(s) = f(t) = f(a)$ then f is constant so $f'(x) = 0$ for all $x \in (a, b)$. Otherwise, $f(t) > f(a)$ and $t \in (a, b)$ or $f(s) < f(a)$ and $s \in (a, b)$. If $s \in (a, b)$ then $(s, f(s))$ is a local minimum for f, so $f'(s) = 0$ by Fermat's Theorem. If $t \in (a, b)$ then $(t, f(t))$ is a local maximum for f, so $f'(t) = 0$ by Fermat's Theorem. The result follows.

□



Theorem 5.16. *The Cauchy Mean Value Theorem. Let f and g be continuous on $[a, b]$ and differentiable on (a, b). Then there is a point $c \in (a, b)$ so that $f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a))$.*

Proof. Let $h(x) = f(x)(g(b) - g(a)) - g(x)(f(b) - f(a))$. Then $h(a) = h(b) = f(a)g(b) - g(a)f(b)$ and $h'(x) = f'(x)(g(b) - g(a)) - g'(x)(f(b) - f(a))$ for all $x \in (a, b)$ and h is continuous on $[a, b]$. By Rolle's Theorem, there is some $c \in (a, b)$ so that $h'(c) = f'(c)(g(b) - g(a)) - g'(c)(f(b) - f(a)) = 0$, so $f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a))$.

□

The Cauchy Mean Value Theorem is also referred to as the Generalized Mean Value Theorem.

Theorem 5.17. *The Mean Value Theorem. Let f be continuous on $[a, b]$ and differentiable on (a, b). Then there is a point $c \in (a, b)$ so that $f'(c)(b - a) = f(b) - f(a)$.*

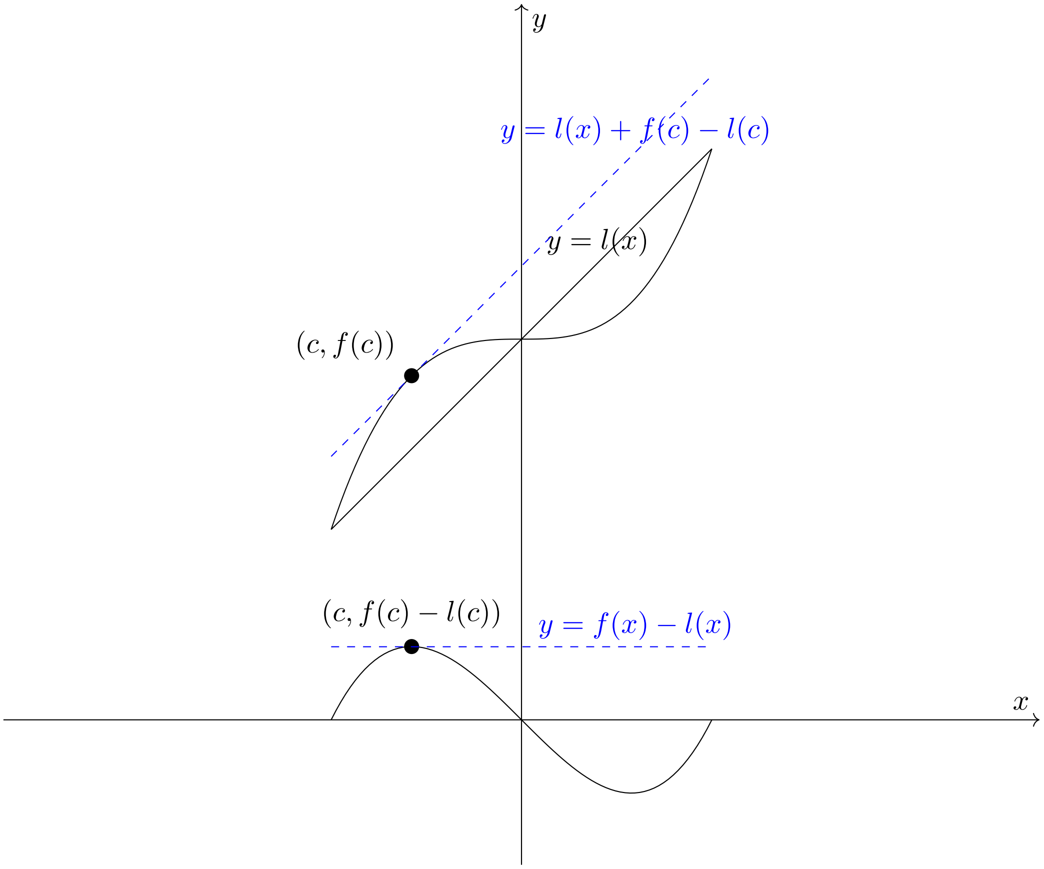
Proof. Setting $g(x) = x$ in the Cauchy Mean Value Theorem, we know that there is a point $c \in (a, b)$ so that $f'(c)(g(b) - g(a)) = g'(c)(f(b) - f(a))$, so $f'(c)(b - a) = f(b) - f(a)$.

□

While the preceding proof is brief, it has less geometric intuitive appeal than the following argument.

Proof. Let $l(x) = \dfrac{f(b) - f(a)}{b - a}(x-a)+f(a)$, the line through $(a, f(a))$ and $(b, f(b))$. Then if $h(x) = f(x) - L(x)$ we note that $h(a) = 0 = h(b)$, so by Rolle's Theorem, for some $c \in (a, b)$ we know that $h'(c) = 0$. Thus, $f'(c) = l'(c) = \dfrac{f(b) - f(a)}{b - a}$. □

Illustration of Mean Value Theorem



The Mean Value Theorem can be used to prove many other theorems. Here are a few of them.

Theorem 5.18. *Let f be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) > 0$ for all $x \in (a, b)$ then f is increasing on $[a, b]$.*

Proof. Let $c, d \in [a, b]$ where $c < d$. Then by the Mean Value Theorem there is a point $t \in (c, d)$ so that $f'(t)(d - c) = f(d) - f(c)$. Since $f'(t) > 0$ it follows that $f(d) - f(c) > 0$ so $f(d) > f(c)$ and f is increasing.

□

Theorem 5.19. *Let f be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) < 0$ for all $x \in (a, b)$ then f is decreasing on $[a, b]$.*

We leave the proof of this theorem as an exercise for the reader.

Theorem 5.20. *Let f, g be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = g'(x)$ for all $x \in (a, b)$ then $f(x) = g(x) + k$ for all $x \in [a, b]$ for some constant k.*

Proof. Let $z \in (a, b]$ and define $h(x) = f(x) - g(x)$. Then by the Mean Value Theorem there is a point $c \in (a, z]$ so that $0 = h'(c)(z - a) = h(z) - h(a)$. Thus, $h(z) = h(a)$, so $f(z) - g(z) = f(a) - g(a)$, so $f(z) = g(z) + (f(a) - g(a))$ for every point $z \in [a, b]$.

□

Note that a special case of the preceding theorem is that if a function has a derivative of zero on an interval then the function is constant.

There are many other theorems can be proven with the Mean Value Theorem. Normally, we guess the Mean Value Theorem might be helpful if we know things about the derivative of a function and we want to know things about how function values at different points compare with one another. Identifying that a theorem can be phrased in those terms is a first step to using the Mean Value Theorem. Here are a couple of additional examples.

Example 5.1. *Let $f'(x) > 5$ on $\mathbb{R}$ and let $f(2) = 4$. Prove $f(4) > 14$.*

Solution. Since f is differentiable on $\mathbb{R}$, f is differentiable and therefore also continuous on $[2, 4]$. Thus, by the Mean Value Theorem, there is a point $c \in (2, 4)$ so that $f'(c)(4 - 2) = f(4) - f(2) = f(4) - 4$. Since $f'(c) > 5$ we know that $2(5) < f(4) - 4$, so $f(4) > 14$. □

Example 5.2. *Prove $|\cos(b) - \cos(a)| \leq |b - a|$ for all real $a < b$.*

Solution. Since $\cos(x)$ is differentiable and therefore continuous everywhere with derivative $(\cos(x))' = -\sin(x)$ by Exercise 5.6, by the Mean Value Theorem we can find some $c \in (a, b)$ so that $-\sin(c)(b - a) = \cos(b) - \cos(a)$, so $|\sin(c)||b - a| = |\cos(b) - \cos(a)|$. Since $|\sin(c)| \leq 1$ it follows that $|b - a| \geq |\cos(b) - \cos(a)|$. □

Example 5.3. *Let $f'(x) < g'(x)$ for all $x \in \mathbb{R}$ and let $f(0) = g(0) = 0$. Then $f(x) < g(x)$ for all $x > 0$ and $f(x) > g(x)$ for all $x < 0$.*

Solution. Let $h(x) = g(x) - f(x)$. Then $h'(x) = g'(x) - f'(x) > 0$ for all $x \in \mathbb{R}$. By the Mean Value Theorem, if $x > 0$ we can choose $c \in (0, x)$ so that $h'(c)(x - 0) = h(x) - h(0) = h(x)$. Since $h'(c) > 0$ we know that $h(x) > 0$ so $g(x) - f(x) > 0$ and $g(x) > f(x)$. Likewise, if $x < 0$ we can find $c \in (x, 0)$ so that $h'(c)(x - 0) = h(x) - h(0) = h(x)$. Since $h'(c) > 0$ we know that $h(x) < 0$ so $g(x) - f(x) < 0$ and $g(x) < f(x)$. □

As an alternate approach, we could have used the fact that h was increasing in the last example.

Theorem 5.21. *Let $f'(x) \neq 0$ on (a, b), and let f be continuous on $[a, b]$. Then f is one to one on $[a, b]$.*

Proof. Let $a \leq x_1 < x_2 \leq b$. Then by the Mean Value Theorem we can find $c \in (a, b)$ so that $f'(c)(x_2 - x_1) = f(x_1) - f(x_2)$. Since $f'(c) \neq 0$ we know that $f(x_1) - f(x_2) \neq 0$, so $f(x_1) \neq f(x_2)$. Thus, f is one to one on $[a, b]$. □

Fermat's Theorem and the preceding theorems about increasing and decreasing functions are also helpful for providing tests to find extrema.

Theorem 5.22. *The First Derivative Test. Let f be differentiable on (a, b) and let $c \in (a, b)$. If $f'(x) \leq 0$ on (a, c) and $f'(x) \geq 0$ on (c, b) then $(c, f(c))$ is a local minimum for f. If $f'(x) \geq 0$ on (a, c) and $f'(x) \leq 0$ on (c, b) then $(c, f(c))$ is a local maximum for f. If $f'(x) > 0$ on (a, c) and $f'(x) > 0$ on (c, b) or $f'(x) < 0$ on (a, c) and $f'(x) < 0$ on (c, b) then $(c, f(c))$ is a saddle point for f.*

Proof. Since (a, b) is open, we can find $\epsilon > 0$ so that $(c - \epsilon, c + \epsilon) \subseteq (a, b)$, so if $f'(x) \leq 0$ on (a, c) and $f'(x) \geq 0$ on (c, b) then f is non-increasing on $[c - \epsilon, c]$ (by Theorem 5.19) and non-decreasing on $[c, c + \epsilon]$ (by Theorem 5.18), which means that if $c - \epsilon < x < c$ then $f(x) \geq f(c)$, and if $c < x < c + \epsilon$ then $f(x) \geq f(c)$, so $(c, f(c))$ is a local minimum for f. Similarly, if $f'(x) \geq 0$ on (a, c) and $f'(x) \leq 0$ on (c, b) then f is non-decreasing on $[c - \epsilon, c]$ and non-increasing on $[c, c + \epsilon]$, which means that if $c - \epsilon < x < c$ then $f(x) \leq f(c)$, and if $c < x < c + \epsilon$ then $f(x) \leq f(c)$, so $(c, f(c))$ is a local maximum for f.

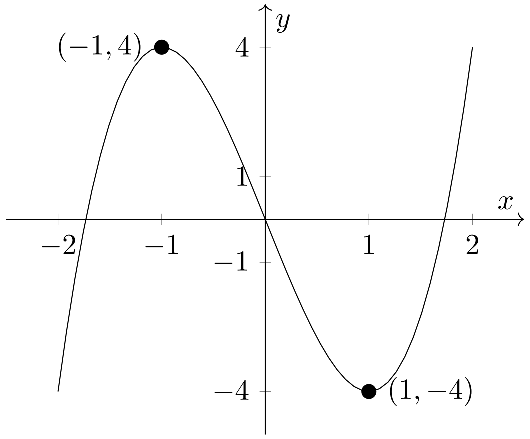
Finally, if $f'(x) > 0$ on (a, c) and $f'(x) > 0$ on (c, b) then f is increasing on both $[a, c]$ and $[c, b]$ which means that $f(c - \frac{\epsilon}{2}) < f(c) < f(c + \frac{\epsilon}{2})$, which means that $(c, f(c))$ is neither a local minimum nor a local maximum, so $(x, f(c))$ is a saddle point. Similarly, $f'(x) < 0$ on (a, c) and $f'(x) < 0$ on (c, b) then f is decreasing on $(c - \epsilon, c + \epsilon)$, so $f(c - \frac{\epsilon}{2}) > f(c) > f(c + \frac{\epsilon}{2})$, which means that $(c, f(c))$ is neither a local minimum nor a local maximum, so $(x, f(c))$ is a saddle point. □

Most of the time the process of finding local extrema using the first derivative test involves first differentiating f and then setting $f'(x) = 0$ and solving. You should also identify the points at which $f'(x)$ is undefined. This gives you a collection of points $x_1, x_2, ..., x_n$ called critical points, where the derivative is either zero or undefined. You then test a value between consecutive critical points (or preceding the first critical point or exceeding the last critical point). It is a consequence of the Intermediate Value Theorem that if $f'(x)$ is continuous on an interval then it cannot change sign without taking on the value zero, so you can normally conclude that the sign of $f'(x)$ is the same for all points between any two critical points. The first derivative test then tells you which of the critical points represent points at which you get extrema and which represent saddle points. This only works for functions that are continuous on an interval and differentiable except at finitely many points.

Example 5.4. *Let $f(x) = 2x^3 - 6x$. Find all local extrema and saddle points.*

Solution: $f'(x) = 6x^2 - 6$ which exists for all values of x, so we set $6x^2 - 6 = 0$ and get $6(x{-}1)(x{+}1) = 0$ so we have $-1, 1$ as critical points. Plugging in values $x = -2, x = 0, x = 2$ to test the sign of $f'(x)$ on each interval gives us $f'(-2) = 18 > 0$, $f'(0) = -6 < 0$ and $f'(2) = 18 > 0$, which means that $(-1, f(-1)) = (-1, 4)$ is a local maximum, whereas $(1, -4)$ is a local minimum.

Graph of f



If $f''(x) > 0$ on interval (a, b) then the rate at which f increases is increasing, or the rate at which f decreases is decreasing. This motivates the second derivative test:

Theorem 5.23. *The Second Derivative Test. Let f be twice differentiable on (a, b) and let $f'(c) = 0$ for some $c \in (a, b)$. Then if $f''(c) < 0$ the point $(c, f(c))$ is a (strict) local maximum and if $f''(c) > 0$ then the point $(c, f(c))$ is a (strict) local minimum.*

Proof. First, assume that $f''(c) = L > 0$. Then we can find a $\delta > 0$ so that if $0 < |x - c| < \delta$ then $x \in (a, b)$ and $|\dfrac{f'(x) - f'(c)}{x - c} - L| < \dfrac{L}{2}$, which means that $\dfrac{f'(x) - 0}{x - c} > \dfrac{L}{2} > 0$. Hence, the sign of $f'(x)$ is the same as the sign of $x - c$, which means that $f'(x) > 0$ if $x \in (c, c + \delta)$ and $f'(x) < 0$ if $x \in (c - \delta, c)$. Thus, $(c, f(c))$ is a local minimum by the first derivative test. If $f''(c) < 0$ then $-f''(c) > 0$ which means that $(c, -f(c))$ is a local minimum for $-f(x)$ and thus $(c, f(c))$ is a local maximum for $f(x)$. □

Example 5.5. *Let $f(x) = \cos(x)$. Find all local extrema for $f(x)$.*

Solution. In this case the extrema are well known (maxima at multiples of 2π and minima at odd multiples of π) but we will use the second derivative test, using $f''(x) = -\sin(x)$ and $f''(x) = -\cos(x)$. The zeroes of $\sin(x)$ are the integer multiples of π, and that $-\cos(n\pi) =$

$1 > 0$ if n is odd and $-\cos(n\pi) = -1 < 0$ if n is even, which means that we have a local maximum of $(n\pi, 1)$ when n is an even integer and a local minimum of $(n\pi, -1)$ when n is an odd integer.

□

L'Hospital's rule has many cases. One can approach from one side or both sides or approach infinity, and the limit of the function derivatives can be infinity, negative infinity or zero and the ratio can approach a real number or an infinite limit. The infinite limits case is a bit awkward and it might be better to prove the simpler case since most problems can be reduced to this case, so we prove only part of the cases for L'Hospital's Rule below and then prove the full version in the Supplementary Materials for those who are interested. Proving the infinity over infinity case is much simpler if we know the limit exists to begin with (and sometimes proofs use that assumption), but without that assumption we have to work harder.

Theorem 5.24. *L'Hospital's Rule (zero over zero case, approaching a real number). Let $f, g : I \to \mathbb{R}$ be differentiable and let $g'(x) \neq 0$ for all $x \in I \setminus \{a\}$ where a is either an element of the open interval I or an end point of I. Let $\lim\limits_{x\to a} f(x) = 0 = \lim\limits_{x\to a} g(x)$ and $\lim\limits_{x\to a} \dfrac{f'(x)}{g'(x)} = L$. Then $\lim\limits_{x\to a} \dfrac{f(x)}{g(x)} = L$.*

Proof. First, define functions F, G by setting $F(x) = f(x)$ if $x \in I \setminus \{a\}$ and set $F(a) = G(a) = 0$. The resulting functions are continuous at a by Exercise 4.3 and Theorem 4.4.

Next, note that since $G'(x) \neq 0$ for all $x \in I \setminus \{a\}$ it must follow that $G(x)$ is one to one on $I \cap [a, \infty)$ and on $I \cap (-\infty, a]$ by Theorem 5.21. Since $G(a) = 0$ we may conclude that $G(x) \neq 0$ for all $x \in I \setminus \{a\}$.

Let $\{x_n\} \to a$, where $\{x_n\} \subset I \setminus \{a\}$. Then by the Cauchy Mean Value Theorem, for each $n \in \mathbb{N}$ we may choose c_n between a and x_n so that $F'(c_n)(G(x_n) - G(a)) = G'(c_n)(F(x_n) - F(a))$. Since $G'(c_n), G(x_n) \neq 0$ and $F(a) = G(a) = 0$, it follows that $\dfrac{F'(c_n)}{G'(c_n)} = \dfrac{f'(c_n)}{g'(c_n)} = \dfrac{f(x_n)}{g(x_n)} = \dfrac{F(x_n)}{G(x_n)}$. We know $0 \leq |c_n - a| < |x_n - a|$ for each natural number n, and $\{|x_n - a|\} \to 0$ by Exercise 3.1. Thus, $\{|c_n - a|\} \to 0$ by the Squeeze Theorem, so $\{c_n\} \to a$. Hence, by the Sequential Characterization of Limits, we know that $\{\dfrac{f'(c_n)}{g'(c_n)}\} \to L$, and thus $\{\dfrac{f(x_n)}{g(x_n)}\} \to L$. From this, it follows that $\lim\limits_{x\to a} \dfrac{f(x)}{g(x)} = L$.

□

Theorem 5.25. *L'Hospital's Rule (zero over zero case, approaching ∞ or $-\infty$). Let $f, g : (a, \infty) \to \mathbb{R}$ be differentiable and let $g'(x) \neq 0$ for all x, where $a > 0$. Let $\lim\limits_{x\to\infty} f(x) = 0 = \lim\limits_{x\to\infty} g(x)$ and $\lim\limits_{x\to\infty} \dfrac{f'(x)}{g'(x)} = L$. Then $\lim\limits_{x\to\infty} \dfrac{f(x)}{g(x)} = L$.*

Furthermore, if we replace ∞ by $-\infty$ and (a, ∞) by $(-\infty, a)$ where $a < 0$, the theorem is still true.

Proof. We begin by addressing the first case, where $f, g : (a, \infty) \to \mathbb{R}$ and $\lim\limits_{x\to\infty} f(x) = 0 = \lim\limits_{x\to\infty} g(x)$. By Theorem 4.14, we know that $\lim\limits_{x\to\infty} f(x) = 0$ if and only if $\lim\limits_{x\to 0^+} f(\dfrac{1}{x}) = 0$.

Likewise, $\lim_{x\to\infty} g(x) = 0$ if and only if $\lim_{x\to 0^+} g(\frac{1}{x}) = 0$. Similarly, $\lim_{x\to\infty} \frac{f'(x)}{g'(x)} = L$ if and only if $\lim_{x\to 0^+} \frac{f'(\frac{1}{x})}{g'(\frac{1}{x})} = L$. By the chain rule, $(f(\frac{1}{x}))' = f'(\frac{1}{x})\frac{-1}{x^2}$ and $(g(\frac{1}{x}))' = g'(\frac{1}{x})\frac{-1}{x^2}$, which means that $\frac{(f(\frac{1}{x}))'}{(g(\frac{1}{x}))'} = \frac{f'(\frac{1}{x})\frac{-1}{x^2}}{g'(\frac{1}{x})\frac{-1}{x^2}} = \frac{f'(\frac{1}{x})}{g'(\frac{1}{x})}$.

Thus, from the the previous form of L'Hospital's rule, since $f(\frac{1}{x}), g(\frac{1}{x}) : (0, \frac{1}{a}) \to \mathbb{R}$ is differentiable and $(g(\frac{1}{x}))'$ is non-zero and $\lim_{x\to 0^+} f(\frac{1}{x}) = 0 = \lim_{x\to 0^+} g(\frac{1}{x})$, and we also know that $\lim_{x\to 0^+} \frac{(f(\frac{1}{x}))'}{(g(\frac{1}{x}))'} = \lim_{x\to 0^+} \frac{f'(\frac{1}{x})}{g'(\frac{1}{x})} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)} = L$, we are able to conclude that $= \lim_{x\to 0^+} \frac{f(\frac{1}{x})}{g(\frac{1}{x})} = L$. Using Theorem 4.14 again gives us that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L$.

The case where we replace ∞ by $-\infty$ is similar except that we replace $x \to 0^+$ by $x \to 0^-$ and replace $(0, \frac{1}{a})$ by $(\frac{1}{a}, 0)$.

□

Theorem 5.26. *Let $f(x)$ be a one to one continuous function on (a, b). Then $f(x)$ is strictly monotone and $f^{-1}(x)$ is continuous and strictly monotone. Furthermore, if $f(x)$ is increasing then $f^{-1}(x)$ is increasing, and if $f(x)$ is decreasing then $f^{-1}(x)$ is decreasing.*

Proof. We first claim that if f is not monotone then there are points $x_1 < x_2 < x_3$ in (a, b) so that either $f(x_1) < f(x_2)$ and $f(x_2) > f(x_3)$ or $f(x_1) > f(x_2)$ and $f(x_2) < f(x_3)$. To prove this claim, suppose that the claim is false. That is, suppose that f is not monotone but for any points $x_1 < x_2 < x_3$ in (a, b), if $f(x_1) < f(x_2)$ then $f(x_2) < f(x_3)$ and if $f(x_1) > f(x_2)$ then $f(x_2) > f(x_3)$. Let $x_1, x_2 \in (a, b)$ with $x_1 < x_2$. Assume $f(x_1) < f(x_2)$. Then if $x_2 < x < y < b$ we know that since $f(x_1) < f(x_2)$ it follows that $f(x_2) < f(x)$, from which we conclude that $f(x) < f(y)$. Likewise, if $x_1 < x < x_2$ then $f(x_1) < f(x)$ since otherwise $f(x) < f(x_2)$ contradicting out supposition that the claim is false, so $f(x_1) < x$ and by the preceding argument then $f(x) < f(x_2)$. Finally, if $a < x < x_1$ then $f(x) < f(x_1)$ since otherwise $f(x) > f(x_1)$ and $f(x_1) < f(x_2)$ contradicting the supposition that the claim is false. It follows that f is increasing on all of (a, b), contradicting the supposition that f is not monotone. Similarly, if $f(x_1) > f(x_2)$ it follows that the function f is decreasing, contradicting to the assumption that f is monotone. Hence, the claim is true.

We now show that f is strictly monotone. Suppse f is not strictly monotone. Then by the claim, there are points $x_1 < x_2 < x_3$ in (a, b) so that either $f(x_1) < f(x_2)$ and $f(x_2) > f(x_3)$ or $f(x_1) > f(x_2)$ and $f(x_2) < f(x_3)$. In the former case, if we choose k between $\max(f(x_1), f(x_3))$ and $f(x_2)$ then by the Intermediate Value Theorem we can find points $c_1 \in (x_1, x_2)$ and $c_2 \in (x_2, x_3)$ so that $f(c_1) = k = f(c_2)$ so f is not one to one. In the latter case the proof is similar, choosing k between $\min(f(x_1), f(x_3))$ and $f(x_2)$. Thus, f is strictly monotone.

Next, assume that f is increasing, and let $y_1 < y_2$ in the range of f. Then there are points $x_1, x_2 \in (a, b)$ so that $f(x_1) = y_1$ and $f(x_2) = y_2$. Suppose $f^{-1}(y_1) \geq f^{-1}(y_2)$. Then since $x_1 \geq x_2$ it follows that $f(x_1) \geq f(x_2)$, so $y_1 \geq y_2$, which is impossible. Similarly, if f is decreasing then f^{-1} must be decreasing.

Finally, to show continuity we will assume that f is strictly increasing, let $y_0 = f(x_0)$ for some $x_0 \in (a,b)$ and let $\epsilon > 0$. Choose $x_1, x_2 \in (a,b)$ so that $x_0 - \epsilon < x_1 < x_0 < x_2 < x_0 + \epsilon$, and let $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Let $\delta = \min(y_0 - y_1, y_2 - y_0)$. Then if $|y - y_0| < \delta$ it follows that $y_1 < y < y_2$ and hence $x_0 - \epsilon < x_1 < f^{-1}(y) < x_2 < x_0 + \epsilon$, so $|f^{-1}(y_0) - f^{-1}(y)| < \epsilon$. Thus, f^{-1} is continuous. If f is decreasing then the argument is similar.

□

Theorem 5.27. *Inverse Function Theorem. Let $f(x)$ be a one to one continuous function on (a,b). If $f'(x_0)$ exists and is non-zero for some $x_0 \in (a,b)$ and $g(x) = f^{-1}(x)$ then $g'(f(x_0)) = \dfrac{1}{f'(x_0)}$.*

Proof. Let $y_0 = f(x_0)$. By Theorem 5.26 we know that g is continuous on $f((a,b))$. Let $\{y_n\}$ be a sequence of points in $f((a,b)) \setminus \{y_0\}$ so that $\{y_n\} \to y_0$.

Then $\{g(y_n)\} \to g(y_0) = x_0$ by the Sequential Characterization of Continuity, and $g(y_n) \neq x_0$ for any $n \in \mathbb{N}$. Since f is differentiable at x_0 we know that $\lim\limits_{x \to x_0} \dfrac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$. Since $f'(x_0) \neq 0$ it follows that $\lim\limits_{x \to x_0} \dfrac{x - x_0}{f(x) - f(x_0)} = \dfrac{1}{f'(x_0)}$, so, by the Sequential Characterization of Limits, $\{\dfrac{g(y_n) - g(y_0)}{f(g(y_n)) - f(g(y_0))}\} \to \dfrac{1}{f'(x_0)}$, where $f(g(y_n)) \neq f(g(y_0))$ for any $n \in \mathbb{N}$ since f is one to one. Since f and g are inverse functions, this means that for any sequence of points $\{y_n\} \subseteq f((a,b)) \setminus \{y_0\}$ so that $\{y_n\} \to y_0$ it is true that $\{\dfrac{g(y_n) - g(y_0)}{f(g(y_n)) - f(g(y_0))}\} = \{\dfrac{g(y_n) - g(y_0)}{y_n - y_0}\} \to \dfrac{1}{f'(x_0)}$. Thus, from the Sequential Characterization of limits again, we conclude that $\lim\limits_{y \to y_0} \dfrac{g(y) - g(y_0)}{y - y_0} = \dfrac{1}{f'(x_0)}$.

□

We used sequences for this proof, but we could have used Theorem 4.11 as well. It might be instructive for the reader to write out a proof of the Inverse Function Theorem using Theorem 4.11.

The Inverse Function Theorem in higher dimensions is quite important and is the key to the Implicit Function Theorem. It is still useful in one variable, primarily for purposes of showing a derivative exists for an inverse function. While the theorem can directly give us the derivative of the inverse function in some cases, in others simply knowing the derivative exists can let us use the chain rule to differentiate the inverse function. When an inverse of a value can be found directly, this theorem can quickly give the derivative of the inverse function based on the derivative of the original function, as shown in the following example.

Example 5.6. *Let $f(x) = x^5 + x^3 + 2x + 1$. Let $g(x) = f^{-1}(x)$. Find $g'(1)$.*

Solution. First, note that $f'(x) = 5x^4 + 3x^2 + 2 > 0$ for all x which means that f is one to one and has an inverse. Since $f(0) = 1$, $g'(1) = \dfrac{1}{f'(0)} = \dfrac{1}{2}$.

□

Exercises:

Exercise 5.1. *Prove Theorem 5.19.*

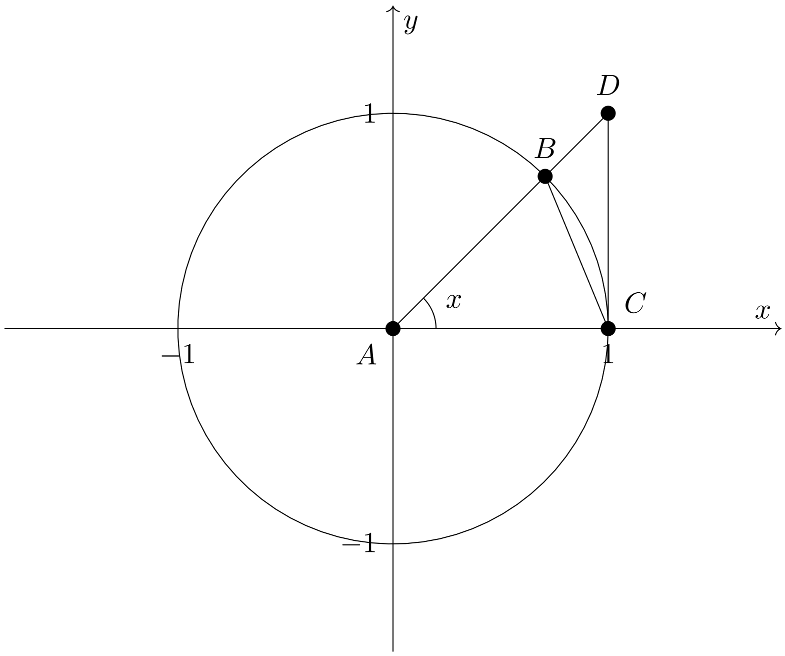
Exercise 5.2. *If $f(x) = x^n$, where $n \in \mathbb{N}$ then $f'(x) = nx^{n-1}$.*

Exercise 5.3. *If $r \in \mathbb{Q}$ then prove that $(x^r)' = rx^{r-1}$ when this expression is defined. In this theorem we use the convention that we replace $0x^{-1}$ by 0 for purposes of this formula even at $x = 0$ and replace x^0 by 1 for all x, even at $x = 0$.*

Exercise 5.4. *If f is differentiable at x_0 and g is differentiable at $f(x_0)$ then $g \circ f$ is defined on an open interval containing x_0.*

Exercise 5.5. *Let $-\frac{\pi}{2} < x < \frac{\pi}{2}$. Let T_1 be the triangle with vertices $(0,0), (1,0)$ and $(\cos(x), \sin(x))$ and let T_2 be the triangle with vertices $(0,0), (1,0)$ and $(1, \tan(x))$. Then the circular sector S of the unit circle between the positive x-axis and the ray based at the origin through the point $(\cos(x), \sin(x))$ contains the triangular disk bounded by T_1 and is contained in the triangular disk T_2. Assuming that the area enclosed by T_1 is understood to be less than or equal to the area within the circular sector S, which is less than or equal to the area enclosed by T_2, and using standard formulas for area of a triangle and a circular sector, as well as standard trigonometric identities, show that $\cos(x) \leq \frac{\sin(x)}{x} \leq 1$, $\lim_{x \to 0} \sin(x) = 0$, $\lim_{x \to 0} \cos(x) = 1$, $\lim_{x \to 0} \frac{\sin(x)}{x} = 1$ and $\lim_{x \to 0} \frac{\cos(x) - 1}{x} = 0$.*

Comparing Areas to Sine



Exercise 5.6. *Using the result that* $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$*, and the sine and cosine sum and difference of angle formulas and pythagorean identities (or the sum to product or other trigonometric identities if preferred, all assumed without proof in this development), show that* $(\sin(x))' = \cos(x)$*,* $(\cos(x))' = -\sin(x)$*,* $(\tan(x))' = \sec^2(x)$*,* $(\csc(x))' = -\csc(x)\cot(x)$*,* $(\sec(x))' = \sec(x)\tan(x)$ *and* $(\cot(x))' = -\csc^2(x)$*.*

Exercise 5.7. *A degree* n *polynomial can have at most* n *real zeroes.*

Exercise 5.8. *Let* $f'(x) > 4$ *for all* $x \in \mathbb{R}$*, and let* $f(0) = 2$*. Then* $f(2) > 10$*.*

Exercise 5.9. *Give an example, with proof, of a function which is continuous but not differentiable.*

Exercise 5.10. *Let* $f : I \to \mathbb{R}$ *be a differentiable function, where* I *is an interval, having the property that* $f'(x)$ *is bounded. Then* f *is uniformly continuous. Furthermore, if* $M > |f'(x)|$ *for all* $x \in I$ *then* $|f(x) - f(y)| < M|x - y|$ *for all* $x, y \in I$ *so that* $x \neq y$*.*

Exercise 5.11. *If* $f'(a) > 0$ *then there is an open interval* (c, d) *containing* a *so that if* $c < x_1 < a < x_2 < d$ *then* $f(x_1) < f(a) < f(x_2)$*.*

Exercise 5.12. *If $f(x)$ is non-decreasing and differentiable on (a,b) then $f'(x) \geq 0$ for all $x \in (a,b)$. Furthermore, if f is non-decreasing and differentiable on $[a,b]$ then $f'(x) \geq 0$ on $[a,b]$.*

Exercise 5.13. *If $f(x)$ is non-increasing and differentiable on (a,b) then $f'(x) \leq 0$ for all $x \in (a,b)$. Furthermore, if f is non-increasing and differentiable on $[a,b]$ then $f'(x) \geq 0$ on $[a,b]$.*

Exercise 5.14. *(a) If f is a differentiable odd function (meaning that $f(x) = -f(-x)$ for all $x \in dom(f)$) then $f'(x)$ is an even function (meaning that $f'(-x) = f'(x)$ for all $x \in dom(f)$).*

(b) If f is an even differentiable function then $f'(x)$ is an odd function.

Exercise 5.15. *Using the Inverse Function Theorem, we can argue that if we restrict* $\sin(x)$ *to the domain $\frac{-\pi}{2} \leq x \leq \frac{\pi}{2}$ then the inverse of this function is differentiable on the interior of its domain by the Inverse Function Theorem. Using this fact (or otherwise), show that* $(\sin^{-1}(x))' = \frac{1}{\sqrt{1-x^2}}$.

Exercise 5.16. *In a similar manner to the preceding theorem, find intervals on which the functions* $\tan(x)$ *and* $\sec(x)$ *could be restricted so that they would be invertible over their restricted domains, and derive formulas for the derivatives of* $\tan^{-1}(x)$ *and* $\sec^{-1}(x)$. *Using the fact that each of these inverse trigonometric functions, when added to its corresponding inverse co-function has a sum of $\frac{\pi}{2}$ (or otherwise) find derivatives for* $\cos^{-1}(x), \csc^{-1}(x)$, *and* $\cot^{-1}(x)$.

Exercise 5.17. *Give an example, with proof, of a function which is differentiable at a point, whose derivative is not differentiable at that point.*

Exercise 5.18. *Give an example, with proof, of a function whose derivative is not bounded which is uniformly continuous.*

Exercise 5.19. *For every non-negative integer n, if $h(x) = f(x)g(x)$, where f and g are n times differentiable, then $h^{(n)}(x) = \sum_{i=0}^{n} \binom{n}{i} f^{(n-i)}(x) g^{(i)}(x)$.*

Exercise 5.20. *Let $\epsilon > 0$ and let $h, g : D \to \mathbb{R}$ be differentiable on $(a-\epsilon, a+\epsilon)$. Let f be a real valued function so that $f(x) = g(x)$ if $a-\epsilon < x < a$ and let $f(x) = h(x)$ if $a < x < a+\epsilon$, where f is continuous at $x = a$. If $\lim_{x \to a^-} g'(x) = k = \lim_{x \to a^+} h'(x)$ then $f'(x) = k$.*

Hints:

Hint to Exercise 5.1. *Prove Theorem 5.19.*

Parallel the proof (or use the result) of Theorem 5.18.

Hint to Exercise 5.2. *If $f(x) = x^n$, where $n \in \mathbb{N}$ then $f'(x) = nx^{n-1}$ (where x^0 is understood to mean* 1*, even at zero).*

Use either induction with the product rule or the Binomial Theorem with the definition of derivative.

Hint to Exercise 5.3. *If $r \in \mathbb{Q}$ then prove that $(x^r)' = rx^{r-1}$ when this expression is defined. In this theorem we use the convention that we replace $0x^{-1}$ by* 0 *for purposes of this formula even at $x = 0$ and replace x^0 by* 1 *for all x, even at $x = 0$.*

Use the Inverse Function Theorem (and probably the chain rule) to differentiate root functions. Use the quotient rule for negative integer powers. Then use the chain rule to differentiate the composition.

Hint to Exercise 5.4. *If f is differentiable at x_0 and g is differentiable at $f(x_0)$ then $g \circ f$ is defined on an open interval containing x_0.*

Remember that to be differentiable at a point implies the function is defined on a neighborhood about that point. Use the continuity of f to find an interval small enough, centered at x_0, so that its image lies within an interval on which g is defined.

Hint to Exercise 5.5. *Let $-\frac{\pi}{2} < x < \frac{\pi}{2}$. Let T_1 be the triangle with vertices $(0,0), (1,0)$ and $(\cos(x), \sin(x))$ and let T_2 be the triangle with vertices $(0,0), (1,0)$ and $(1, \tan(x))$. Then the circular sector S of the unit circle between the positive x-axis and the ray based at the origin through the point $(\cos(x), \sin(x))$ contains the triangular disk bounded by T_1 and is contained in the triangular disk T_2. Assuming that the area enclosed by T_1 is understood to be less than or equal to the area within the circular sector S, which is less than or equal to the area enclosed by T_2, and using standard formulas for area of a triangle and a circular sector, as well as standard trigonometric identities, show that $\cos(x) \leq \frac{\sin(x)}{x} \leq 1$, $\lim_{x \to 0} \sin(x) = 0$, $\lim_{x \to 0} \cos(x) = 1$, $\lim_{x \to 0} \frac{\sin(x)}{x} = 1$ and $\lim_{x \to 0} \frac{\cos(x) - 1}{x} = 0$.*

Use the Squeeze Theorem and the identity $\tan(x) = \frac{\sin(x)}{\cos(x)}$ and the fact that $\sin(x)$ is an odd function and $\cos(x)$ is an even function.

Hint to Exercise 5.6. *Using the result that* $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$, *and the sine and cosine sum and difference of angle formulas and pythagorean identities (or the sum to product or other trigonometric identities if preferred, all assumed without proof in this development), show that* $(\sin(x))' = \cos(x)$, $(\cos(x))' = -\sin(x)$, $(\tan(x))' = \sec^2(x)$, $(\csc(x))' = -\csc(x)\cot(x)$, $(\sec(x))' = \sec(x)\tan(x)$ *and* $(\cot(x))' = -\csc^2(x)$.

Use identities for the derivative of sine and cosine (pythagorean or sum to product). The other function derivatives follow from the quotient rule.

Hint to Exercise 5.7. *A degree n polynomial can have at most n real zeroes.*

Use induction and Rolle's Theorem.

Hint to Exercise 5.8. *Let $f'(x) > 4$ for all $x \in \mathbb{R}$, and let $f(0) = 2$. Then $f(2) > 10$.*

Use the Mean Value Theorem.

Hint to Exercise 5.9. *Give an example, with proof, of a function which is continuous but not differentiable.*

Try to think of a function that forms a corner or cusp or has a derivative approaching infinity at a point (probably at zero).

Hint to Exercise 5.10. *Let $f : I \to \mathbb{R}$ be a differentiable function, where I is an interval, having the property that $f'(x)$ is bounded. Then f is uniformly continuous. Furthermore, if $M > |f'(x)|$ for all $x \in I$ then $|f(x) - f(y)| < M|x - y|$ for all $x, y \in I$ so that $x \neq y$.*

Use the Mean Value Theorem.

Hint to Exercise 5.11. *If $f'(a) > 0$ then there is an open interval (c, d) containing a so that if $c < x_1 < a < x_2 < d$ then $f(x_1) < f(a) < f(x_2)$.*

Choose a short enough open interval containing a so that the difference quotient $\frac{f(x) - f(a)}{x - a} > 0$ on that interval.

Hint to Exercise 5.12. *If $f(x)$ is non-decreasing and differentiable on (a, b) then $f'(x) \geq 0$ for all $x \in (a, b)$. Furthermore, if f is non-decreasing and differentiable on $[a, b]$ then $f'(x) \geq 0$ on $[a, b]$.*

Use the Comparison Theorem.

Hint to Exercise 5.13. *If $f(x)$ is non-increasing and differentiable on (a, b) then $f'(x) \leq 0$ for all $x \in (a, b)$. Furthermore, if f is non-increasing and differentiable on $[a, b]$ then $f'(x) \geq 0$ on $[a, b]$.*

Use the Comparison Theorem.

Hint to Exercise 5.14. *(a) If f is a differentiable odd function (meaning that $f(x) = -f(-x)$ for all $x \in dom(f)$) then $f'(x)$ is an even function (meaning that $f'(-x) = f'(x)$ for all $x \in dom(f)$).*

(b) If f is an even differentiable function then $f'(x)$ is an odd function.
Use either the definition of derivative or the chain rule.

Hint to Exercise 5.15. *Using the Inverse Function Theorem, we can argue that if we restrict $\sin(x)$ to the domain $\frac{-\pi}{2} \leq x \leq \frac{\pi}{2}$ then the inverse of this function is differentiable on the interior of its domain by the Inverse Function Theorem. Using this fact (or otherwise), show that $(\sin^{-1}(x))' = \frac{1}{\sqrt{1-x^2}}$.*

Start with $y = \sin^{-1}(x)$ so $\sin(y) = x$. Then use the Inverse Function Theorem to conclude y is differentiable, and differentiate both sides using the chain rule.

Hint to Exercise 5.16. *In a similar manner to the preceding theorem, find intervals on which the functions $\tan(x)$ and $\sec(x)$ could be restricted so that they would be invertible over their restricted domains, and derive formulas for the derivatives of $\tan^{-1}(x)$ and $\sec^{-1}(x)$. Using the fact that each of these inverse trigonometric functions, when added to its corresponding inverse co-function has a sum of $\frac{\pi}{2}$ (or otherwise) find derivatives for $\cos^{-1}(x), \csc^{-1}(x)$, and $\cot^{-1}(x)$.*

Normally, the convention is to restrict $\cos(x)$ to $[0, \pi]$, $\cot(x)$ to $(0, \pi)$, and $\tan(x)$ to $(-\frac{\pi}{2}.\frac{\pi}{2})$, restrict $\sec(x)$ to $[0, \frac{\pi}{2}) \cup [\pi, \frac{3\pi}{2})$ and $\csc(x)$ to $(0, \frac{\pi}{2}] \cup (\pi, \frac{3\pi}{2}]$. Then use the same process as outlined in the previous exercise.

Hint to Exercise 5.17. *Give an example, with proof, of a function which is differentiable at a point, whose derivative is not differentiable at that point.*

Try to get the second derivative to either approach infinity or simply not exist. Think of a function that is not differentiable at zero which has an antiderivative.

Hint to Exercise 5.18. *Give an example, with proof, of a function whose derivative is not bounded which is uniformly continuous.*

Think of a function which is continuous on a closed interval whose slope approaches infinity (consider a rounded or rapidly oscillating function approaching a vertical tangent).

Hint to Exercise 5.19. *For every non-negative integer n, if $h(x) = f(x)g(x)$, where f and g are n times differentiable, then $h^{(n)}(x) = \sum_{i=0}^{n} \binom{n}{i} f^{(n-i)}(x)g^{(i)}(x)$.*

Model the proof after the argument for the Binomial Theorem using the product rule.

Hint to Exercise 5.20. *Let $\epsilon > 0$ and let $h, g : D \to \mathbb{R}$ be differentiable on $(a-\epsilon, a+\epsilon)$. Let f be a real valued function so that $f(x) = g(x)$ if $a - \epsilon < x < a$ and let $f(x) = h(x)$ if $a < x < a + \epsilon$, where f is continuous at $x = a$. If $\lim_{x \to a^-} g'(x) = k = \lim_{x \to a^+} h'(x)$ then $f'(x) = k$.*

Take one sided limits and use Theorem 4.15.

Solutions:

Solution to Exercise 5.1. *Prove Theorem 5.19.*

Proof. Let $c, d \in [a, b]$, where $c < d$. By the Mean Value Theorem, there is some point $t \in (c, d)$ so that $f'(t)(d - c) = f(d) - f(c)$. Since $f'(t) < 0$ and $(d - c) > 0$, it follows that $f'(t)(d - c) < 0$, so $f(d) - f(c) < 0$, which means that $f(d) < f(c)$. Hence, f is decreasing on $[a, b]$. □

Solution to Exercise 5.2. *If $f(x) = x^n$, where $n \in \mathbb{N}$ then $f'(x) = nx^{n-1}$ (where x^0 is understood to mean 1, even at zero).*

Proof. We can use the Binomial Theorem or induction. We will use induction. We have already shown that the derivative of $y = x$ is $1 = 1x^0$. We assume that the formula is true for a natural number k. Then $(x^{k+1})' = (xx^k)' = (1)(x^k) + x(kx^{k-1}) = (k + 1)x^k$ by the product rule and the induction hypothesis. The result follows by induction. □

Solution to Exercise 5.3. *If $r \in \mathbb{Q}$ then prove that $(x^r)' = rx^{r-1}$ when this expression is defined. In this theorem we use the convention that we replace $0x^{-1}$ by 0 for purposes of this formula even at $x = 0$ and replace x^0 by 1 for all x, even at $x = 0$.*

Proof. Let $r = \frac{p}{q}$ where p is an integer and q is a natural number. First, note that if p is zero the derivative is zero. If p is a negative integer then $-p$ is a natural number, so by the preceding argument $(x^p)' = (\frac{1}{x^{-p}})' = \frac{0 - -px^{-p-1}}{x^{-2p}}$ using the quotient rule. This simplifies to px^{p-1}. Next, using the Inverse Function Theorem, we know that since $x^{\frac{1}{q}}$ is the inverse of x^q, the function $y = x^{\frac{1}{q}}$ is differentiable. Hence $y^q = x$ and by the chain rule $qy^{q-1}y' = 1$ so $y' = \frac{1}{q}x^{\frac{1}{q}-1}$. Finally, again using the chain rule, we have that if $y = x^{\frac{p}{q}} = (x^{\frac{1}{q}})^p$ then $y' = p((x^{\frac{1}{q}})^{p-1})(x^{\frac{1}{q}})' = p((x^{\frac{1}{q}})^{p-1})\frac{1}{q}x^{\frac{1}{q}-1} = \frac{p}{q}x^{\frac{p}{q}-1}$ as desired.

□

Solution to Exercise 5.4. *If f is differentiable at x_0 and g is differentiable at $f(x_0)$ then $g \circ f$ is defined on an open interval containing x_0.*

Proof. Since g is differentiable at $f(x_0)$ there is an $\epsilon_g > 0$ so that $(f(x_0) - \epsilon_g, f(x_0) + \epsilon_g) \subset dom(g)$. Since f is differentiable at x_0 there is an $\epsilon_f > 0$ so that $(x_0 - \epsilon_f, x_0 + \epsilon_f) \subset dom(f)$. Since f is differentiable at x_0, by a f is continuous at x_0 by Theorem 5.3. Thus, we can find $\delta_1 > 0$ so that if $|x_0 - x| < \delta_1$ and $x \in dom(f)$ then $|f(x) - f(x_0)| < \epsilon_g$. Thus, if $|x - x_0| < \delta = \min\{\epsilon_f, \delta_1\}$ then$x \in dom(f)$ and $|f(x) - f(x_0)| < \epsilon_g$, so $(x_0 - \delta, x_0 + \delta) \subset dom(g \circ f)$. □

Solution to Exercise 5.5. *Let* $-\frac{\pi}{2} < x < \frac{\pi}{2}$*. Let* T_1 *be the triangle with vertices* $(0,0),(1,0)$ *and* $(\cos(x),\sin(x))$ *and let* T_2 *be the triangle with vertices* $(0,0),(1,0)$ *and* $(1,\tan(x))$*. Then the circular sector* S *of the unit circle between the positive* x*-axis and the ray based at the origin through the point* $(\cos(x),\sin(x))$ *contains the triangular disk bounded by* T_1 *and is contained in the triangular disk* T_2*. Assuming that the area enclosed by* T_1 *is understood to be less than or equal to the area within the circular sector* S*, which is less than or equal to the area enclosed by* T_2*, and using standard formulas for area of a triangle and a circular sector, as well as standard trigonometric identities, show that* $\cos(x) \leq \frac{\sin(x)}{x} \leq 1$*,* $\lim_{x\to 0}\sin(x) = 0$*,* $\lim_{x\to 0}\cos(x) = 1$*,* $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$ *and* $\lim_{x\to 0}\frac{\cos(x)-1}{x} = 0$*.*

Proof. The area of triangle $\triangle ABC$ is $\frac{\sin(x)}{2}$, which is less than the area of the sector of the circle ABC, which is $\frac{x}{2}$, which is less than the area of triangle $\triangle ADC$, which is $\frac{\tan(x)}{2}$. Hence, we have $\sin(x) < x < \tan(x)$ for small positive x (and $-x < \sin(x) < 0$ for small negative x). It follows that $\frac{\sin(x)}{x} < 1$. Since $x < \frac{\sin(x)}{\cos(x)}$, we know that $\cos(x) < \frac{\sin(x)}{x}$. Next, since $0 \leq |\sin(x)| \leq |x|$ for $-\frac{-\pi}{2} < x < \frac{\pi}{2}$ from this picture, we conclude from Exercise 4.1 (or just the Squeeze Theorem) that $\lim_{x\to 0}\sin(x) = 0$ since $\lim_{x\to 0}|x| = 0$. Next, note that $\cos(x) = \sqrt{1-\sin^2(x)}$ for all x since $\sqrt{x}$ is defined for all values $1-\sin^2(x)$, so, since $f(x) = \sqrt{x}$ is continuous, we have that $\lim_{x\to 0}\cos(x) = 1$ by Theorem 4.11. Since $\cos(x) < \frac{\sin(x)}{x} < 1$ for $0 < x < \frac{\pi}{2}$, and $\lim_{x\to 0}\cos(x) = 1$ and $\lim_{x\to 0} 1 = 1$, it follows from the Squeeze Theorem that $\lim_{x\to 0^+}\frac{\sin(x)}{x} = 1$. If x is negative then both $\sin(x)$ and $\tan(x)$ are negated, so $\frac{\sin(x)}{x}$ is still between $\cos(x)$ and 1, which means that $\lim_{x\to 0^-}\frac{\sin(x)}{x} = 1$, so $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$.

Since $\lim_{x\to 0}\frac{\cos(x)-1}{x} = \lim_{x\to 0}\frac{(\cos(x)-1)(\cos(x)+1)}{x(\cos(x)+1)} = \lim_{x\to 0}\frac{\cos^2(x)-1}{x(\cos(x)+1)}$ $= \lim_{x\to 0}\frac{-\sin^2(x)}{x(\cos(x)+1)} = \lim_{x\to 0}\frac{\sin(x)}{x}\frac{-\sin(x)}{\cos(x)+1} = (1)(\frac{0}{2}) = 0.$ □

Solution to Exercise 5.6. *Using the result that* $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$*, and the sine and cosine sum and difference of angle formulas and pythagorean identities (or the sum to product or other trigonometric identities if preferred, all assumed without proof in this development), show that* $(\sin(x))' = \cos(x)$*,* $(\cos(x))' = -\sin(x)$*,* $(\tan(x))' = \sec^2(x)$*,* $(\csc(x))' = -\csc(x)\cot(x)$*,* $(\sec(x))' = \sec(x)\tan(x)$ *and* $(\cot(x))' = -\csc^2(x)$*.*

Proof. First, note that $\lim_{h\to 0}\frac{\cos(h)-1}{h} = \lim_{h\to 0}\frac{(\cos(h)-1)(\cos(h)+1)}{h(\cos(h)+1)} = \lim_{h\to 0}\frac{(\cos^2(h)-1)}{h(\cos(h)+1)} =$ $\lim_{h\to 0}\frac{\sin^2(h)}{h(\cos(h)+1)} = \lim_{h\to 0}\frac{\sin(h)}{h}\frac{\sin(h)}{(\cos(h)+1)} = 0.$

Thus, $(\cos(x))' = \lim_{h\to 0} \frac{\cos(x+h) - \cos(x)}{h} = \lim_{h\to 0} \frac{\cos(x)\cos(h) - \sin(x)\sin(h) - \cos(x)}{h} = \lim_{h\to 0} -\sin(x)\frac{\sin(h)}{h} + \cos(x)\frac{\cos(h) - 1}{h} = -\sin(x)(1) + 0$ by Theorem 4.6, so $(\cos(x))' = -\sin(x)$. Likewise, $(\sin(x))' = \lim_{h\to 0} \frac{\sin(x+h) - \sin(x)}{h} = \lim_{h\to 0} \frac{\sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x)}{h} = \lim_{h\to 0} \sin(x)\frac{\cos(h) - 1}{h} + \cos(x)\frac{\sin(h)}{h} = \cos(x)$.

From this we can use the quotient rule to obtain the remaining derivatives:

$(\tan(x))' = (\frac{\sin(x)}{\cos(x)})' = \frac{\cos(x)\cos(x) + \sin(x)\sin(x)}{\cos^2(x)} = \sec^2(x)$.

$(\cot(x))' = (\frac{\cos(x)}{\sin(x)})' = \frac{-\sin(x)\sin(x) - \cos(x)\cos(x)}{\sin^2(x)} = -\csc^2(x)$

$(\sec(x))' = (\frac{1}{\cos(x)})' = \frac{0 + \sin(x)}{\cos^2(x)} = \sec(x)\tan(x)$

$(\csc(x))' = (\frac{1}{\sin(x)})' = \frac{0 - \cos(x)}{\sin^2(x)} = -\csc(x)\cot(x)$

□

Solution to Exercise 5.7. *A degree n polynomial can have at most n real zeroes.*

Proof. Proceed by induction. A degree one polynomial has form $P(x) = ax+b$ where $a \neq 0$. Thus, there is exactly one solution $x = \frac{-b}{a}$. Assume that a degree k polynomial can have at most k real zeroes. Let $P(x) = a_{k+1}x^{k+1} + ... + a_1x + a_0$ be a degree $k+1$ polynomial. Then $P'(x) = (k+1)x^k + ... + a_1$ is a degree k polynomial and has no more than k real zeroes by the induction hypothesis. Suppose P has $k+2$ zeroes $x_1 < x_2 < ...x_{k+2}$. Then by Rolle's Theorem, there is a point $c_i \in (x_i, x_{i+1})$ so that $P'(c_i) = 0$ for each $1 \leq i \leq k+1$, which means that P' has $k+1$ zeroes, contradicting the induction hypothesis. The result follows by induction. □

Solution to Exercise 5.8. *Let $f'(x) > 4$ for all $x \in \mathbb{R}$, and let $f(0) = 2$. Then $f(2) > 10$.*

Proof. Since f is differentiable at all real numbers, it is also continuous at all real numbers, so by the Mean Value Theorem there is a point $c \in (0,2)$ so that $f'(c)(2-0) = f(2)-f(0) = f(2) - 2$. Since $f'(c) > 4$ we know that $f(2) - 2 > 8$ and hence $f(2) > 10$. □

Solution to Exercise 5.9. *Give an example, with proof, of a function which is continuous but not differentiable.*

Proof. The most common example is $|x|$. We know that $f(x) = x$ and $f(x) = -x$ are both continuous by earlier theorems, so $|x|$ is continuous at every point except zero. However, $\lim_{x\to 0^-} |x| = \lim_{x\to 0^-} -x = |0| = \lim_{x\to 0^+} x = \lim_{x\to 0^+} |x|$, so it follows that $|x|$ is also continuous at 0.

On the other hand $\lim_{x\to 0^-} \frac{|x| - 0}{x - 0} = \lim_{x\to 0^-} \frac{-x}{x} = -1$ whereas $\lim_{x\to 0^+} \frac{|x| - 0}{x - 0} = \lim_{x\to 0^+} \frac{x}{x} = 1$. Thus, $|x|$ is not differentiable at 0.

□

Solution to Exercise 5.10. *Let $f : I \to \mathbb{R}$ be a differentiable function, where I is an interval, having the property that $f'(x)$ is bounded. Then f is uniformly continuous. Furthermore, if $M > |f'(x)|$ for all $x \in I$ then $|f(x) - f(y)| < M|x - y|$ for all $x, y \in I$ so that $x \neq y$.*

Proof. Choose $M > 0$ so that $|f'(x)| < M$ for all $x \in I$ and let $\epsilon > 0$. Then setting $\delta = \dfrac{\epsilon}{M}$ we know that if $|x - y| < \delta$ for $x, y \in I$, then by the Mean Value Theorem there is a point $c \in (x, y)$ so that $f'(c)(y - x) = f(y) - f(x)$ which means that $|f'(c)||y - x| = |f(y) - f(x)|$ and therefore $|f(y) - f(x)| < \dfrac{\epsilon}{M} M = \epsilon$. Thus, f is uniformly continuous.

More generally, for any $x < y$ in I we can pick c between x and y so that $|f'(c)||y - x| = |f(y) - f(x)|$ by the Mean Value Theorem, which means that $|f(x) - f(y)| < M|x - y|$.

□

Solution to Exercise 5.11. *If $f'(a) > 0$ then there is an open interval (c, d) containing a so that if $c < x_1 < a < x_2 < d$ then $f(x_1) < f(a) < f(x_2)$.*

Proof. First, by the definition of differentiable we can choose $\epsilon_1 > 0$ so that $(a - \epsilon_1, a + \epsilon_1) \subset dom(f)$. Next, since $f'(a) > 0$ we can find $0 < \delta < \epsilon_1$ so that if $|x - x_0| < \delta$ then $|\dfrac{f(x) - f(a)}{x - a} - f'(a)| < \dfrac{f'(a)}{2}$, and thus $\dfrac{f(x) - f(a)}{x - a} > \dfrac{f'(a)}{2} > 0$. If $a - \delta < x_1 < a < x_2 < a + \delta$ we know that $x_1 - a < 0$ and $x_2 - a > 0$, so it follows that $f(x_1) < f(a)$ and $f(a) < f(x_2)$.

□

Solution to Exercise 5.12. *If $f(x)$ is non-decreasing and differentiable on (a, b) then $f'(x) \geq 0$ for all $x \in (a, b)$ Furthermore, if f is non-decreasing and differentiable on $[a, b]$ then $f'(x) \geq 0$ on $[a, b]$.*

Proof. Fix $x \in (a, b)$. Since f is non-decreasing on (a, b), we know that if $|h|$ is small enough so that $(x - |h|, x + |h|) \subset (a, b)$ then $\dfrac{f(x + h) - f(x)}{h} \geq 0$. Thus, by the Comparison Theorem for limits we know that $\lim\limits_{h \to 0} \dfrac{f(x + h) - f(x)}{h} = f'(x) \geq 0$. If f is also differentiable at a and b then similarly we have $\lim\limits_{h \to 0^+} \dfrac{f(a + h) - f(a)}{h} = f'(a) \geq 0$ and $\lim\limits_{h \to 0^-} \dfrac{f(b + h) - f(b)}{h} = f'(b) \geq 0$.

□

Solution to Exercise 5.13. *If $f(x)$ is non-increasing and differentiable on (a, b) then $f'(x) \leq 0$ for all $x \in (a, b)$. Furthermore, if f is non-increasing and differentiable on $[a, b]$ then $f'(x) \leq 0$ on $[a, b]$.*

Proof. Fix $x \in (a,b)$. Since f is non-increasing on (a,b), we know that if $|h|$ is small enough so that $(x-|h|, x+|h|) \subset (a,b)$ then $\dfrac{f(x+h)-f(x)}{h} \leq 0$. Thus, by the Comparison Theorem for limits we know that $\lim\limits_{h\to 0} \dfrac{f(x+h)-f(x)}{h} = f'(x) \leq 0$. If f is also differentiable at a and b then similarly we have $\lim\limits_{h\to 0^+} \dfrac{f(a+h)-f(a)}{h} = f'(a) \leq 0$ and $\lim\limits_{h\to 0^-} \dfrac{f(b+h)-f(b)}{h} = f'(b) \leq 0$.

□

Solution to Exercise 5.14. *(a) If f is a differentiable odd function (meaning that $f(x) = -f(-x)$ for all $x \in dom(f)$) then $f'(x)$ is an even function (meaning that $f'(-x) = f'(x)$ for all $x \in dom(f)$).*

(b) If f is an even differentiable function then $f'(x)$ is an odd function.

Proof. (a) Using the Chain Rule, $(f(-x))' = -f'(-x)$. On the other hand, $f(-x) = -f(x)$ and $(-f(x))' = -f'(x)$. Thus, $f'(-x) = f'(x)$ and so f' is even.

(b) Using the Chain Rule, $(f(-x))' = -f'(-x)$. On the other hand, $f(-x) = f(x)$ and $(f(x))' = f'(x)$. Thus, $-f'(-x) = f'(x)$ and so f' is odd.

□

Solution to Exercise 5.15. *Using the Inverse Function Theorem, we can argue that if we restrict $\sin(x)$ to the domain $\frac{-\pi}{2} \leq x \leq \frac{\pi}{2}$ then the inverse of this function is differentiable on the interior of its domain by the Inverse Function Theorem. Using this fact (or otherwise), show that $(\sin^{-1}(x))' = \dfrac{1}{\sqrt{1-x^2}}$.*

Proof. As mentioned, the Inverse Function Theorem guarantees that if $y = \sin^{-1}(x)$ then y' exists, so by the Chain Rule we have $\sin(y) = x$ so $\cos(y)y' = 1$, which means $y' = \dfrac{1}{\cos(y)} = \dfrac{1}{\sqrt{1-\sin^2(y)}} = \dfrac{1}{\sqrt{1-x^2}}$.

□

Solution to Exercise 5.16. *In a similar manner to the preceding theorem, find intervals on which the functions $\tan(x)$ and $\sec(x)$ could be restricted so that they would be invertible over their restricted domains, and derive formulas for the derivatives of $\tan^{-1}(x)$ and $\sec^{-1}(x)$. Using the fact that each of these inverse trigonometric functions, when added to its corresponding inverse co-function has a sum of $\frac{\pi}{2}$ (or otherwise) find derivatives for $\cos^{-1}(x), \csc^{-1}(x)$, and $\cot^{-1}(x)$.*

Proof. Normally, the convention is to restrict $\cos(x)$ to $[0,\pi]$, $\cot(x)$ to $(0,\pi)$, and $\tan(x)$ to $(-\frac{\pi}{2}.\frac{\pi}{2})$. There are multiple conventions for $\sec(x)$ and $\csc(x)$ but it seems to be easiest if we restrict $\sec(x)$ to $[0,\frac{\pi}{2})\cup[\pi,\frac{3\pi}{2})$ and $\csc(x)$ to $(0,\frac{\pi}{2}]\cup(\pi,\frac{3\pi}{2}]$ because this makes derivatives simpler.

Using the same methods described above, the Inverse Function Theorem guarantees that each of these inverse functions is differentiable, so we proceed as we did for $\sin^{-1}(x)$.

If $y=\tan^{-1}(x)$ then $\tan(y)=x$ so $y'\sec^2(y)=1$, which means that $y'=\dfrac{1}{\sec^2(y)}=\dfrac{1}{1+\tan^2(y)}=\dfrac{1}{1+x^2}$.

If $y=\sec^{-1}(x)$ then $\sec(y)=x$ so $y'\sec(y)\tan(y)=1$, which means that $y'=\dfrac{1}{\sec(y)\tan(y)}=\dfrac{1}{\sec(y)\sqrt{\tan^2(y)-1}}=\dfrac{1}{x\sqrt{x^2-1}}$.

Since $\cos^{-1}(x)=\frac{\pi}{2}-\sin^{-1}(x)$ we know that $(\cos^{-1}(x))'=\dfrac{-1}{\sqrt{1-x^2}}$.

Since $\cot^{-1}(x)=\frac{\pi}{2}-\tan^{-1}(x)$ we know that $(\cot^{-1}(x))'=\dfrac{-1}{1+x^2}$.

Since $\csc^{-1}(x)=\frac{\pi}{2}-\sec^{-1}(x)$ we know that $(\csc^{-1}(x))'=\dfrac{-1}{x\sqrt{x^2-1}}$.

□

Solution to Exercise 5.17. *Give an example, with proof, of a function which is differentiable at a point, whose derivative is not differentiable at that point.*

Proof. Let $f(x)=x^2\sin(\frac{1}{x})$ if $x\neq 0$ and let $f(0)=0$. Differentiation at every point other than zero can be conducted using the Chain Rule and Product Rule to give $f'(x)=2x\sin(\frac{1}{x})-\cos(\frac{1}{x})$. However, at $x=0$ we have $f'(x)=\lim_{x\to 0}\dfrac{x^2\sin(\frac{1}{x})-0}{x-0}=\lim_{x\to 0}x\sin(\frac{1}{x})=0$ by Exercise 4.16. Thus, $f'(x)=2x\sin(\frac{1}{x})-\cos(\frac{1}{x})$ if $x\neq 0$ and $f'(0)=0$. Since $\{\frac{1}{2n\pi}\}\to 0$ and $\{f'(\frac{1}{2n\pi})\}\to -1$, it follows that f' is not continuous at 0 and therefore not differentiable at 0.

□

Solution to Exercise 5.18. *Give an example, with proof, of a function whose derivative is not bounded which is uniformly continuous.*

Proof. Let $f(x)=x\sin(\frac{1}{x})$ on $(0,1]$. Then f is differentiable and by the Chain Rule, Quotient Rule and Product Rule we know that $f'(x)=\sin(\frac{1}{x})-\dfrac{\cos(\frac{1}{x})}{x}$. Note that $\{f'(\frac{1}{2n\pi})\}=\{-2n\pi\}$ which is unbounded, so f' is unbounded.

To see that f is uniformly continuous, let $0<\epsilon<1$. By Theorem 4.18, since f is continuous f is also uniformly continuous on the closed interval $[\frac{\epsilon}{4},1]$. Thus, there is a

$\delta_1 > 0$ so that if $|x-y| < \delta$ then $|f(x)-f(y)| < \epsilon$ if $x, y \in [\frac{\epsilon}{4}, 1]$. We set $\delta = \min\{\frac{\epsilon}{4}, \delta_1\}$. Let $|x-y| < \delta$ with $y > x$. We know that $x\sin(\frac{1}{x})$ is between (or equal to) x and $-x$ for all $x \in (0,1]$ since $-1 \leq \sin(\frac{1}{x}) \leq 1$ for all x. Thus, the largest value f takes on in the interval $[x,y]$ cannot exceed y and the smallest value is at least $-y$, so $|f(y)-f(x)| \leq 2y$. Since $y - x < \frac{\epsilon}{4}$ it follows that if $y \geq \frac{\epsilon}{2}$ then $x \geq \frac{\epsilon}{4}$, so since $|x-y| < \delta_1$ we know that $|f(x)-f(y)| < \epsilon$. Otherwise $y < \frac{\epsilon}{2}$ which means that $|f(y)-f(x)| \leq 2y < \epsilon$. Hence, f is uniformly continuous.

□

Solution to Exercise 5.19. *For every non-negative integer n, if $h(x) = f(x)g(x)$, where f and g are n times differentiable, then $h^{(n)}(x) = \sum_{i=0}^{n} \binom{n}{i} f^{(n-i)}(x)g^{(i)}(x)$.*

Proof. First, the result is immediate if $n = 0$. Assume the result when $n = k$. Then $h^{(k+1)}(x) = (\sum_{i=0}^{k} \binom{k}{i} f^{(k-i)}(x)g^{(i)}(x))'$, which, by the product rule, is $\sum_{i=0}^{k} \binom{k}{i} f^{(k+1-i)}(x)g^{(i)}(x) + f^{(k-i)}(x)g^{(i+1)}(x)$, which can be written as $\binom{k}{0} f^{(k+1)}(x)g(x) + \binom{k}{k} f(x)g^{(k+1)}(x) + \sum_{i=1}^{k}(\binom{k}{i} + \binom{k-1}{i})f^{(k+1-i)}(x)g^{(i)}(x)$. We know $\binom{k}{0} = \binom{k+1}{0} = 1$ and $\binom{k}{k} = \binom{k+1}{k+1} = 1$, and $(\binom{k}{i} + \binom{k-1}{i}) = \binom{k+1}{i}$ by Theorem 2.7. Hence, $h^{(k+1)}(x) = \sum_{i=0}^{k+1} \binom{k+1}{i} f^{(k+1-i)}(x)g^{(i)}(x)$, as desired. The result follows by induction. □

Solution to Exercise 5.20. *Let $\epsilon > 0$ and let $h, g : D \to \mathbb{R}$ be differentiable on $(a-\epsilon, a+\epsilon)$. Let f be a real valued function so that $f(x) = g(x)$ if $a - \epsilon < x < a$ and let $f(x) = h(x)$ if $a < x < a + \epsilon$, where f is continuous at $x = a$. If $\lim_{x\to a^-} g'(x) = k = \lim_{x\to a^+} h'(x)$ then $f'(x) = k$.*

Proof. We know h, g are continuous at a since they are differentiable at a. Since f, g and h are continuous at a we know that $\lim_{x\to a^-} g(x) = g(a) = f(x) = \lim_{x\to a^-} f(x)$ since $f(x) = g(x)$ for $x \in (a-\epsilon, a)$. Likewise, $\lim_{x\to a^+} h(x) = h(a) = f(a)$. It follows that $\lim_{x\to a^-} \frac{f(x)-f(a)}{x-a} = \lim_{x\to a^-} \frac{g(x)-g(a)}{x-a} = g'(a) = k$ and $\lim_{x\to a^+} \frac{f(x)-f(a)}{x-a} = \lim_{x\to a^+} \frac{h(x)-h(a)}{x-a} = h'(a) = k$. Since $g'(a) = h'(a)$ the left and right limits are equal and $\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = f'(a) = k$. □

Chapter 6

Integration

There are different approaches to doing integration theorems. Three popular methods are first, using upper and lower sums to get upper and lower integrals to identify the integral when an integral exists, second, using limits of Riemann sums with markings that can be taken within subintervals induced by partitions and defining the integral to the be the limit of such sums if this limit is unaffected by the markings, and third, using a theorem of Lebesgue that a bounded function is integrable if and only if the set of discontinuities of the function has Lebesgue measure zero. In the third case, we get a powerful tool for determining integrability, but we still have to use another method to actually find the integral. We will use a development based on the first two ways of describing integrability, and address the third technique in the Supplementary Materials section. We may give multiple proofs of some results when different approaches seem to have advantages. In the case where we use Lebesgue's characterization of Riemann integrability for the argument we will also give an argument using either upper and lower sums or Riemann sums (so no theorem in this section will rely in using the Supplementary Materials).

Definition 32

Let $f : [a,b] \to \mathbb{R}$ be bounded. A finite subset $P = \{x_0, x_1, ..., x_n\} \subset [a,b]$, understood to be listed in order with $x_0 = a < x_1 < x_2 < ... < x_n = b$ is called a *partition* of $[a,b]$. If P and Q are partitions of $[a,b]$ and $P \subseteq Q$ then we say that Q is a *refinement* of P (or that Q refines P). A collection of points $T = \{x_1^*, x_2^*, ..., x_n^*\}$ so that $x_i^* \in [x_{i-1}, x_i]$ for each $i \in \{1,2,3,...,n\}$ is called a *marking* of P, and we denote the *Riemann sum* with this marking by $S_T(f,P) = \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1})$. If we let $M_i = \sup_{x \in [x_{i-1},x_i]} f(x)$ and $m_i = \inf_{x \in [x_{i-1},x_i]} f(x)$ then we say that $U(f,P) = \sum_{i=1}^{n} M_i(x_i - x_{i-1})$ is the *upper sum* of f with respect to P, and $L(f,P) = \sum_{i=1}^{n} m_i(x_i - x_{i-1})$ is the *lower sum* of f with respect to P. If the partition is understood then we sometimes use the notation M_i, m_i for suprema and infima of function values on induced subintervals of the partition without declaring them to be thus defined in an argument. We call $|P| = \max(x_1 - x_0, x_2 - x_1, ..., x_n - x_{n-1})$ the *mesh* of the partition P. The

upper integral of f on $[a,b]$ is denoted $(U)\int_a^b f$, and is the infimum of all upper sums of f on $[a,b]$. The *lower integral* of f on $[a,b]$ is denoted $(L)\int_a^b f$, and is the supremum of all lower sums of f on $[a,b]$. If the upper and lower integrals are equal then we say that f is *integrable* on $[a,b]$ and that the *integral* of f over $[a,b]$ is $\int_a^b f = (L)\int_a^b f = (U)\int_a^b f$.
Let $\mathbb{R}^2 = \mathbb{R}\times\mathbb{R}$, the set of ordered pairs (a,b), where a and b are real numbers.
If $f(x) \geq g(x)$ for all $a \leq x \leq b$ then we say that $\int_a^b f(x) - g(x)dx$ is the *area* between the curves $y = f(x)$ and $y = g(x)$ in $\mathbb{R}^2$.

In some cases it may be useful to distinguish between infima and suprema of different functions on the same subintervals induced by a partition or on different partitions.

Definition 33

Let $M_i^f(P)$ denote the supremum on $[x_{i-1}, x_i]$, the ith subinterval induced by the partition P for the function f and $m_i^f(P)$ to refer to the infimum on $[x_{i-1}, x_i]$ for the function f. If the partition is understood we may just write M_i^f, m_i^f, and if the function is understood we may write $M_i(P), m_i(P)$ without the superscript.
It is sometimes convenient to take suprema or infima over all partitions (or other types of sets) without specifying the set of all partitions, as long as the interval over which the partitions is taken is understood. We use the notation $\sup_P f(P)$ and $\inf_P f(P)$ to denote the supremum and infimum, respectively, of the function $f(P)$ over all possible partitions P the interval in question (or sets P which are understood to be an element of a certain collection).

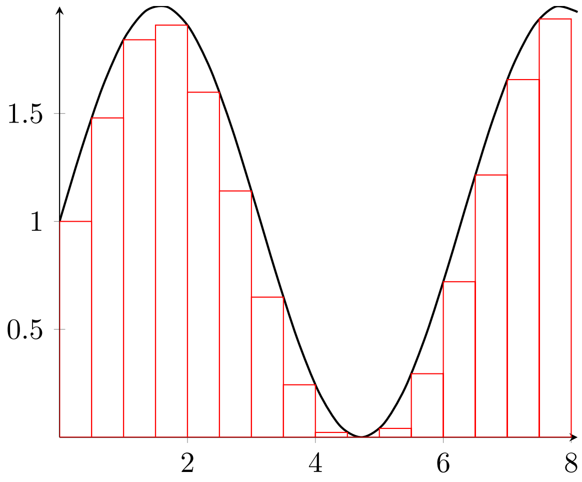
Not all functions are integrable. Unbounded functions are not integrable, but many bounded functions are also not integrable. Here is an example of such a function.

Example 6.1. *Give an example of a bounded function which is not integrable.*

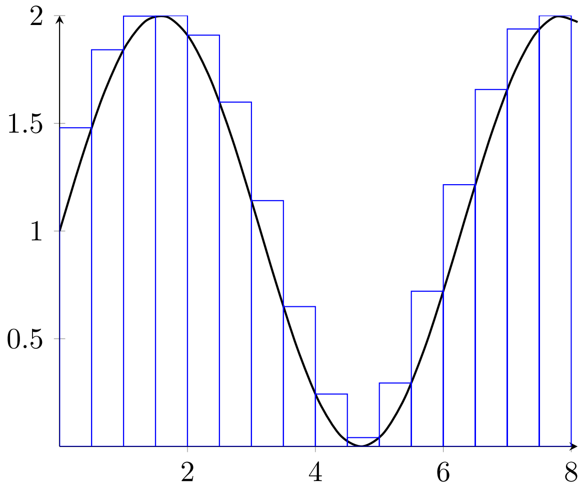
Solution. Let $f(x) = 0$ if $x \in \mathbb{Q}$ and let $f(x) = 1$ otherwise. Let $P = \{x_0, x_1, x_2, ..., x_n\}$ be a partition of $[0,1]$. Since every $[x_{i-1}, x_i]$ subinterval induced by P contains both rational and irrational numbers, $U(f,P) = \sum_{i=1}^{n} M_i(x_i - x_{i-1}) = \sum_{i=1}^{n}(1)(x_i - x_{i-1}) = 1$, whereas $L(f,P) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}) = \sum_{i=1}^{n}(0)(x_i - x_{i-1}) = 0$. Thus, the upper integral $(U)\int_0^1 f = 1$ and the lower integral $(L)\int_0^1 f = 0$. Since these are not equal, f is not integrable on $[0,1]$.

□

Below is an illustration of a lower sum. The area beneath the red rectangles is $L(f, P)$ if $P = \{0, \frac{1}{2}, 1, \frac{3}{2}, ..., \frac{15}{2}, 8\}$ and $f(x) = \sin(x) + 1$. Since the function is continuous, the infimum on each subinterval induced by the partition is just the minimum value of the function on the subinterval. Multiplying the subinterval length by the minimum value of the function on the subinterval gives the area enclosed by the rectangles shown, which is less than the area under the curve.



Below is a picture of the upper sum of the same function with the same partition. The area enclosed by the blue rectangles is the upper sum. Note that the area under the curve would be between the upper and lower sums.

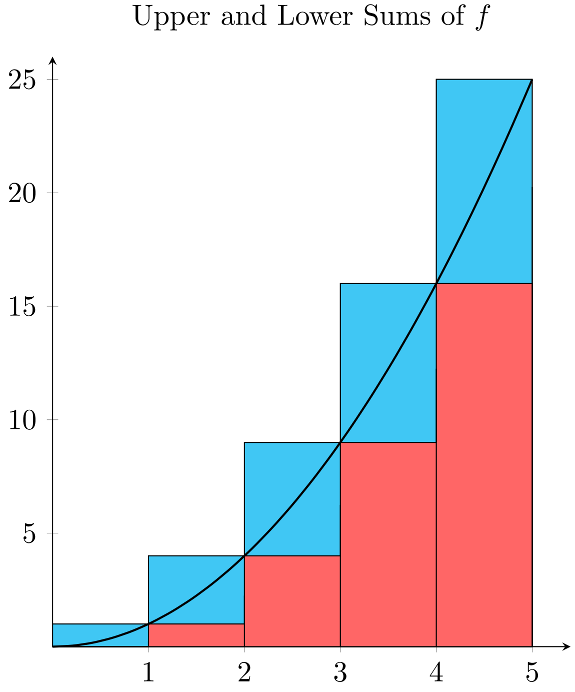


Theorem 6.1. *Let $f : [a, b] \to \mathbb{R}$ be bounded. Let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$. Then for any marking T of P, $L(f, P) \leq S_T(f, P) \leq U(f, P)$.*

Proof. Let $T = \{x_1^*, x_2^*, ..., x_n^*\}$ be a marking for P. Then for each $i \leq n$ we know that $m_i \leq f(x_i^*) \leq M_i$, so $\displaystyle\sum_{i=1}^{n} m_i(x_i - x_{i-1}) \leq \sum_{i=1}^{n} f(x_i^*)(x_i - x_{i-1}) \leq \sum_{i=1}^{n} M_i(x_i - x_{i-1})$.

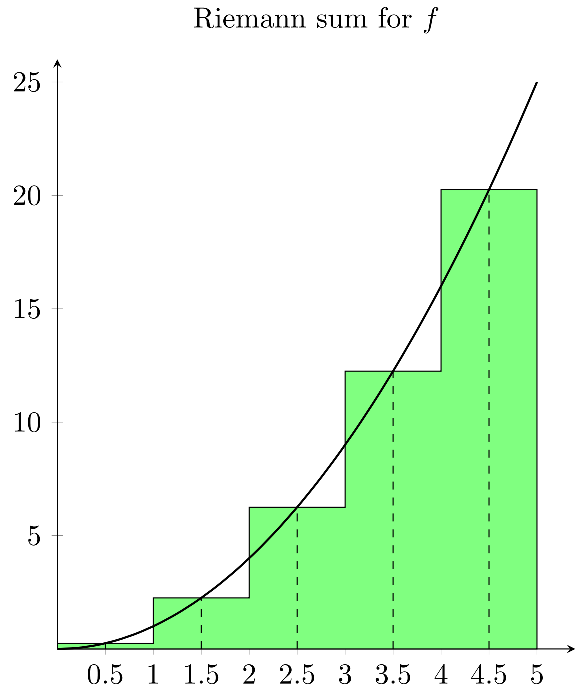
□

In the following example, let $P = \{0, 1, 2, 3, 4, 5\}$ be a partition of $[0, 5]$. We sketch a graph of $U(f, P)$ and $L(f, P)$ for $f(x) = x^2$. In this case, both upper and lower sums are sketched in the same picture. we have taken fewer subdivisions in this example, and the upper and lower sums do not look very close to the area beneath the curve (which the integral since the function is non-negative). In general, we will find that for integrable functions, if we can make the mesh of the partition small enough (usually by taking more partition points that are evenly spaced in the interval) then we can force the resulting upper, lower and Riemann sums to be as close as we wish to the integral, whereas such sums tend not to be close to the integral when the number of subdivisions is small.



The area in blue is the difference between the upper and lower sums. The area in red is the lower sum. The sum of the two shown areas is the upper sum. Note that the upper sum for a partition is an area that is at least equal to the area under the curve, the lower sum for the partition is no more than the area under the curve, and that integral functions are bounded functions so that the difference between the upper and lower sums (the blue area) can be made arbitrarily small.

A Riemann sum is at least as large as the lower sum and no larger than the upper sum for a given partition. It is formed by taking function values at points in each subinterval and multiplying those function values by the subinterval lengths and adding the result. This corresponds to an area like that shown below. We use the same function and partition as the preceding example for this picture. We will use the midpoints of the intervals for the marking, however.



Riemann sum for f

Theorem 6.2. *Let $f : [a,b] \to \mathbb{R}$ be bounded. Let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a,b]$, and let Q be a refinement of P. Then $L(f,P) \leq L(f,Q) \leq U(f,Q) \leq U(f,P)$.*

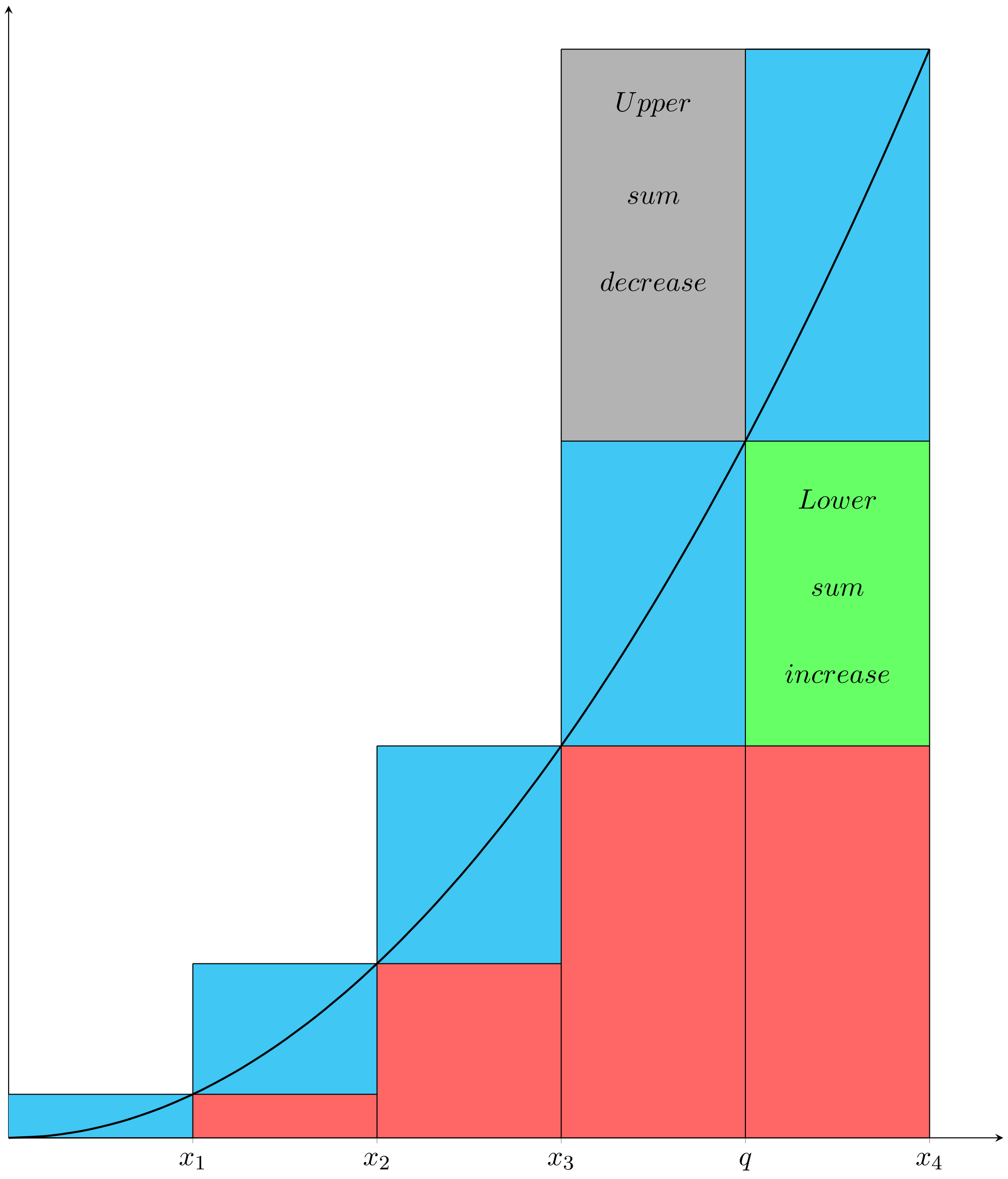
Proof. We first prove the result is true when $Q = P \cup \{q\}$, where $q \in (x_{i-1}, x_i)$, and $M_i = \sup\limits_{x\in[x_{i-1},x_i]} f(x)$. Then $U(f,P) - U(f,Q) = M_i(x_i - x_{i-1}) - \sup\limits_{x\in[x_{i-1},q]} f(x)(q - x_{i-1}) - \sup\limits_{x\in[q,x_i]} f(x)(x_i - q) \geq 0$ since $M_i \geq \max(\sup\limits_{x\in[x_{i-1},q]} f(x), \sup\limits_{x\in[q,x_i]} f(x))$ by Exercise 1.17. Similarly, $L(f,P) - L(f,Q) = m_i(x_i - x_{i-1}) - \inf\limits_{x\in[x_{i-1},q]} f(x)(q - x_{i-1}) - \inf\limits_{x\in[q,x_i]} f(x)(x_i - q) \leq 0$ since $m_i \leq \min(\inf\limits_{x\in[x_{i-1},q]} f(x), \inf\limits_{x\in[q,x_i]} f(x))$. Thus, $L(f,P) \leq L(f,Q) \leq U(f,Q) \leq U(f,P)$.

Next, let $Q = P \cup \{q_1, q_2, ..., q_m\}$. Then by the first part of the argument, it follows that $L(f,P) \leq L(f, P \cup \{q_1\}) \leq L(f, P \cup \{q_1, q_2\}) \leq ... \leq L(f,Q) \leq U(f,Q) \leq U(f, P \cup \{q_1, q_2, ..., q_{m-1}\}) \leq U(f, P \cup \{q_1, q_2, ..., q_{m-2}\}) \leq ... \leq U(f,P)$.

□

The following figure illustrates why a refinement adding a single point causes the upper sum to decrease and the lower sum to increase.

Refinement With Added Point q



Theorem 6.3. *Let $f : [a, b] \to \mathbb{R}$ be bounded. Let P and Q be partitions of $[a, b]$. Then $L(f, P) \leq U(f, Q)$. Furthermore, $(L)\int_a^b f \leq (U)\int_a^b f$.*

Proof. By Theorem 6.2 we know that $L(f, P) \leq L(f, P \cup Q) \leq U(f, P \cup Q) \leq U(f, Q)$. Since for any partition Q of $[a, b]$ it is true that $U(f, Q) \geq L(f, P)$ for each partition P of $[a, b]$, we know that $U(f, Q) \geq (L)\int_a^b f$, so $(L)\int_a^b f$ a lower bound for all upper sums $U(f, Q)$ of $[a, b]$, which means that $(L)\int_a^b f \leq (U)\int_a^b f$.

□

Theorem 6.4. *Let $f : [a,b] \to \mathbb{R}$ be bounded. Then f is integrable if and only if for every $\epsilon > 0$ there is a partition R of $[a,b]$ so that $U(f,R) - L(f,R) < \epsilon$.*

Proof. First, assume that f is integrable. By the Approximation Property, we can find partitions P and Q of $[a,b]$ so that $U(f,P) < (U)\displaystyle\int_a^b f + \frac{\epsilon}{2}$ and $L(f,Q) > (U)\displaystyle\int_a^b f - \frac{\epsilon}{2}$, which means that $\displaystyle\int_a^b f - \frac{\epsilon}{2} < L(f,Q) \leq L(f,P \cup Q) \leq U(f,P \cup Q) \leq U(f,Q) < \int_a^b f + \frac{\epsilon}{2}$. Thus, it follows that $U(f,P \cup Q) - L(f,P \cup Q) < \epsilon$.

Next, assume that for every $\epsilon > 0$ there is a partition R of $[a,b]$ so that $U(f,R) - L(f,R) < \epsilon$. Let $\epsilon > 0$. Choose R so that $U(f,R) - L(f,R) < \epsilon$. But then by Theorem 6.3 we know that $L(f,R) \leq (L)\displaystyle\int_a^b f \leq (U)\int_a^b f \leq U(f,R)$, so $0 \leq (U)\displaystyle\int_a^b f - (L)\int_a^b f < \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $(L)\displaystyle\int_a^b f = (U)\int_a^b f$, so f is integrable.

□

Theorem 6.5. *Let $f : [a,b] \to \mathbb{R}$ be bounded, and let $\epsilon > 0$ and let $P = \{x_0, ..., x_n\}$ be a partition of $[a,b]$. Then there are markings T and R of P so that $U(f,P) - S_T(f,P) < \epsilon$ and $S_R(f,P) - L(f,P) < \epsilon$.*

Proof. Since $M_i = \displaystyle\sup_{x \in [x_{i-1},x_i]} f(x)$ and $m_i = \displaystyle\inf_{x \in [x_{i-1},x_i]} f(x)$, by the Approximation Property we can find points $r_i^*, t_i^* \in [x_{i-1}, x_i]$ for each $i \in \{1,2,3,...,n\}$ so that $M_i - f(t_i^*) < \dfrac{\epsilon}{b-a}$ and $f(r_i^*) - m_i < \dfrac{\epsilon}{b-a}$ to obtain markings $T = \{t_1^*, t_2^*, t_3^*, ...t_n^*\}$ and $R = \{r_1^*, r_2^*, r_3^*, ...r_n^*\}$ so that $U(f,P) - S_T(f,P) = \displaystyle\sum_{i=1}^{n}(M_i - f(t_i^*))(x_i - x_{i-1}) < \frac{\epsilon}{b-a}\sum_{i=1}^{n}(x_i - x_{i-1}) = \frac{\epsilon}{b-a}(b-a) = \epsilon$ and $S_R(f,P) - L(f,P) = \displaystyle\sum_{i=1}^{n}(f(r_i^*) - m_i)(x_i - x_{i-1}) < \frac{\epsilon}{b-a}\sum_{i=1}^{n}(x_i - x_{i-1}) = \frac{\epsilon}{b-a}(b-a) = \epsilon$.

□

Theorem 6.6. *Let $f : [a,b] \to \mathbb{R}$ be integrable. Then for every $\epsilon > 0$ there is a number $\delta > 0$ so that if Q is a partition of $[a,b]$ with $|Q| < \delta$ then $U(f,Q) - L(f,Q) < \epsilon$.*

Proof. Since f is integrable, we know that we can find $M > 0$ so that $|f(x)| < M$ for all $x \in [a,b]$ and we can find a partition $P = \{x_0, ..., x_n\}$ so that $U(f,P) - L(f,P) < \dfrac{\epsilon}{2}$. Let $\delta = \dfrac{\epsilon}{4nM}$, and let $Q = \{q_0, ..., q_m\}$ be a partition with $|Q| < \delta$. Then $U(f,Q) - L(f,Q) =$

$$\sum_{\{j \in \mathbb{N} | [q_{j-1},q_j] \subseteq [x_{i-1},x_i] \text{ for some } i \leq n\}} (M_j^f(Q) - m_j^f(Q))(q_j - q_{j-1})$$

$$+ \sum_{\{j \in \mathbb{N} | x_i \in (q_{j-1},q_j) \text{ for some } i \leq n\}} (M_j^f(Q) - m_j^f(Q))(q_j - q_{j-1}).$$

For each $i \in \{1,2,...,n\}$ we know that if $[q_{j-1},q_j] \subseteq [x_{i-1},x_i]$ then $m_i^f(P) \leq m_j^f(Q) \leq M_j^f(Q) \leq M_i^f(P)$ by Exercise 1.17, so $\displaystyle\sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq[x_{i-1},x_i]} (M_j^f(Q) - m_j^f(Q))(q_j - q_{j-1}) \leq (M_i^f(P) - m_i^f(P))(\sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq[x_{i-1},x_i]} (q_j - q_{j-1})) \leq (M_i^f(P) - m_i^f(P))(x_i - x_{i-1})$. Thus, $\displaystyle\sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq[x_{i-1},x_i] \text{ for some } i\leq n\}} (M_j^f(Q) - m_j^f(Q))(q_j - q_{j-1}) \leq U(f,P) - L(f,P) < \frac{\epsilon}{2}$.

Since $|Q| < \dfrac{\epsilon}{4nM}$ and there are at most $n-1$ integers j so that $x_i \in (q_{j-1},q_j)$ for some $x_i \in P$, it follows that $\displaystyle\sum_{\{j\in\mathbb{N}|x_i\in(q_{j-1},q_j) \text{ for some } x_i\in P\}} (M_j^f(Q) - m_j^f(Q))(q_j - q_{j-1}) \leq 2M(n-1)(\frac{\epsilon}{4nM}) < \frac{\epsilon}{2}$. Thus, $U(f,Q) - L(f,Q) < \epsilon$.

□

Theorem 6.7. *Let $f : [a,b] \to \mathbb{R}$ be bounded and let $\{P_n\}$ be a sequence of partitions of $[a,b]$ so that $\{|P_n|\} \to 0$. Then $\displaystyle\int_a^b f(x)$ exists if and only if there is a number I so that for any choice of markings T_i of P_i, for each $i \in \mathbb{N}$, the sequence $\{S_{T_n}(f,P_n)\} \to I$, in which case $\displaystyle\int_a^b f(x) = I$ and, for every $\epsilon > 0$, there is a $k \in \mathbb{N}$ so that if $n \geq k$ then $|S_{T_n}(f,P_n) - I| < \epsilon$ regardless of marking T_n.*

Proof. Let $\{P_n\}$ be a sequence of partitions of $[a,b]$ so that $\{|P_n|\} \to 0$ and let $\epsilon > 0$.

First, assume $\displaystyle\int_a^b f(x)dx = I$. By Theorem 6.6 we can find a $\delta > 0$ so that if $|P| < \delta$ then $U(f,P) - L(f,P) < \epsilon$. For some $k \in \mathbb{N}$, if $n \geq k$ then $|P_n| < \delta$. Since $I, S_{T_n}(f,P_n) \in [L(f,P_n), U(f,P_n)]$ (regardless of marking T_n), it follows that $|S_{T_n}(f,P_n) - I| < \epsilon$, so $\{S_{T_n}(f,P_n)\} \to I$.

Next, assume that for every choice of markings T_n of P_n, $\{S_{T_n}(f,P_n)\} \to I$. By Theorem 6.5, for each $n \in \mathbb{N}$ we can choose markings U_n and L_n of P_n so that $U(f,P_n) - S_{U_n}(f,P_n) < \dfrac{\epsilon}{4}$ and $S_{L_n}(f,P_n) - L(f,P_n) < \dfrac{\epsilon}{4}$. Choose $N \in \mathbb{N}$ so that if $n \geq N$ then $|S_{U_n}(f,P_n) - I| < \dfrac{\epsilon}{4}$ and $|S_{L_n}(f,P_n) - I| < \dfrac{\epsilon}{4}$. Then $U(f,P_N) - L(f,P_N) \leq |U(f,P_N) - S_{U_N}(f,P_N)| + |S_{U_N}(f,P_N) - I| + |I - S_{L_N}(f,P_N)| + |S_{L_N}(f,P_N) - L(f,P_N)| < \epsilon$, so f is integrable. Also, $|U(f,P_N) - I| \leq |U(f,P_N) - S_{U_N}(f,P_N)| + |S_{U_N}(f,P_N) - I| < \dfrac{\epsilon}{2}$ and $\displaystyle\int_a^b f(x)dx \in [U(f,P_N), L(f,P_N)]$, so $|U(f,P_N) - \displaystyle\int_a^b f(x)dx| < \epsilon$. Hence, $|\displaystyle\int_a^b f(x)dx - I| \leq |U(f,P_N) - \displaystyle\int_a^b f(x)dx| + |U(f,P_N) - I| < \dfrac{3\epsilon}{2}$. Since this is true for every $\epsilon > 0$ it must follow that $\displaystyle\int_a^b f(x)dx = I$.

□

This establishes the characterization of integral in terms of Riemann sums and upper and lower sums. To see the characterization of integrability in terms of a set of discontinuities having Lebesgue measure zero, see the Supplementary Materials.

Theorem 6.8. *If $f : [a, b] \to \mathbb{R}$ is continuous then f is integrable.*

Proof. Since f is continuous on the closed and bounded interval $[a, b]$, from Theorem 4.18, we know that f is uniformly continuous. Let $\epsilon > 0$. Choose $\delta > 0$ so that if $|x - y| < \delta$ then $|f(x) - f(y)| < \frac{\epsilon}{b-a}$. Let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$ with $|P| < \delta$. By the Extreme Value Theorem there are points $s_i, t_i \in [x_{i-1}, x_i]$ for each positive integer $i \leq n$ so that $f(s_i) \leq f(x) \leq f(t_i)$ for each $x \in [x_{i-1}, x_i]$. Hence, $U(f, P) - L(f, P) = \sum_{i=0}^{n}(f(t_i) - f(s_i))(x_i - x_{i-1}) < \frac{\epsilon}{b-a}\sum_{i=0}^{n}(x_i - x_{i-1}) = \epsilon$. Thus, f is integrable. □

Alternate proof using theorems from the Supplementary Materials:

Proof. Since f is continuous, the set of points in the domain of f at which f is not continuous has Lebesgue measure zero, so by Theorem 7.73, f is integrable. □

Theorem 6.9. *If f and g are integrable on $[a, b]$ and $s, t \in \mathbb{R}$ then $\int_a^b sf + tg = s\int_a^b f + t\int_a^b g$.*

Proof. For any partition $P = \{x_0, x_1, ..., x_k\}$ of $[a, b]$ and marking $T = \{x_1^*, x_2^*, ..., x_k^*\}$ of P, the Riemann sum $S_T(sf + tg) = \sum_{i=1}^{k} sf(x_i^*) + tg(x_i^*) = s\sum_{i=1}^{k} f(x_i^*) + t\sum_{i=1}^{k} g(x_i^*) = sS_T(f, P) + tS_T(g, P)$.

Let $\{P_n\}$ be a sequence of partitions of $[a, b]$ so that $\{|P_n|\} \to 0$ and let T_n be a marking of P_n for each $n \in \mathbb{N}$. By Theorem 6.7, we know that $\{S_{T_n}(f, P_n)\} \to \int_a^b f$ and $\{S_{T_n}(g, P_n)\} \to \int_a^b g$. Thus, by the product and sum theorems for limits of sequences, $\{S_{T_n}(sf + tg, P_n)\} = \{sS_{T_n}(f, P_n) + tS_{T_n}(g, P_n)\} \to s\int_a^b f + t\int_a^b g$. Thus, by Theorem 6.7, we know that $\int_a^b sf + tg = s\int_a^b f + t\int_a^b g$.

□

Theorem 6.10. *Let $f : [a, b] \to [c, d]$, where f is integrable on $[a, b]$ and g is continuous on $[c, d]$. Then $g \circ f$ is integrable.*

Proof. Let $\epsilon > 0$. Choose M so that $|g(x)| < M$ for all $x \in [c, d]$. By Theorem 4.18, we know that g is uniformly continuous on $[c, d]$. We choose $\delta > 0$ so that if $|x - y| < \delta$ and $x, y \in [c, d]$ then $|g(x) - g(y)| < \frac{\epsilon}{2(b-a)}$. Choose a partition $P = \{x_0, ..., x_n\}$ so that

$U(f,P) - L(f,P) < \dfrac{\delta\epsilon}{4M}$. Then $L = \displaystyle\sum_{\{i\in\mathbb{N}|M_i^f - m_i^f \geq \delta\}} (x_i - x_{i-1}) < \dfrac{\epsilon}{4M}$ since otherwise $U(f,P) - L(f,P) \geq L\delta \geq \dfrac{\delta\epsilon}{4M}$.

If $M_i^f - m_i^f < \delta$ then let $\gamma > 0$. By the Approximation Property we can find $x, y \in [x_{i-1}, x_i]$ so that $M_i^{g\circ f} - g(f(x)) < \dfrac{\gamma}{2}$ and $g(f(y)) - m_i^{g\circ f} < \dfrac{\gamma}{2}$, so $M_i^{g\circ f} - m_i^{g\circ f} < g(f(x)) - g(f(y)) + \gamma < \dfrac{\epsilon}{2(b-a)} + \gamma$ (since $|f(x) - f(y)| < \delta$). Since this is true for all $\gamma > 0$, it follows that $M_i^{g\circ f} - m_i^{g\circ f} \leq \dfrac{\epsilon}{2(b-a)}$. Thus, it follows that $U(g\circ f, P) - L(g\circ f, P)$ $= \displaystyle\sum_{\{i\in\mathbb{N}|M_i^f - m_i^f \geq \delta\}} (M_i^{g\circ f} - m_i^{g\circ f})(x_i - x_{i-1}) + \sum_{\{i\in\mathbb{N}|M_i^f - m_i^f < \delta\}} (M_i^{g\circ f} - m_i^{g\circ f})(x_i - x_{i-1}) <$ $2M(\dfrac{\epsilon}{4M}) + (b-a)\dfrac{\epsilon}{2(b-a)} = \epsilon$. □

Alternate proof using theorems from the Supplementary Materials:

Proof. Let $E_f = \{x \in [a,b] | f$ is not continuous at $x\}$, and $E_{g\circ f} = \{x \in [a,b] | g\circ f$ is not continuous at $x\}$. By Theorem 7.73 we know $\lambda(E_f) = 0$, and we know that $E_{g\circ f} \subseteq E_f$ by Theorem 4.12, so $\lambda(E_{g\circ f}) = 0$ by Theorem 7.60, so the result follows from Theorem 7.73. □

Theorem 6.11. *Let $f, g : [a,b] \to \mathbb{R}$ be integrable. Then fg is integrable.*

Proof. First, we know that $g(x) = x^2$ is integrable because it is continuous. Hence, by Theorem 6.10, we know that the square of an integrable function is integrable. Since $fg = \dfrac{1}{4}((f+g)^2 - (f-g)^2)$, it follows that fg is a sum of integrable functions and is integrable by Theorem 6.9.

□

Alternate proof using theorems from the Supplementary Materials:

Proof. Let $E_f = \{x \in [a,b] | f$ is not continuous at $x\}$, and $E_g = \{x \in [a,b] | g$ is not continuous at $x\}$ and let $E_{fg} = \{x \in [a,b] | fg$ is not continuous at $x\}$. By Theorem 7.73, $\lambda(E_f) = 0 = \lambda(E_g)$. By Theorem 4.7, we know that $E_{fg} \subseteq E_f \cup E_g$. By Theorem 7.62, we know that $\lambda(E_f \cup E_g) = 0$, so by Theorem 7.60 it follows that $\lambda(E_{fg}) = 0$. Thus, fg is integrable by Theorem 7.73. □

Theorem 6.12. *First Mean Value Theorem for Integrals. Let $f, g : [a,b] \to \mathbb{R}$ where f is continuous and g is integrable, and let $g(x) \geq 0$ for all $x \in [a,b]$. Then there is a point $c \in [a,b]$ so that $\int_a^b f(x)g(x)dx = f(c)\int_a^b g(x)dx$.*

Proof. First, integrability of fg follows from Theorem 6.11. By the Extreme Value Theorem there are points $s, t \in [a, b]$ so that $f(s) \leq f(x) \leq f(t)$ for all $x \in [a, b]$. Hence, since $g(x) \geq 0$ it follows that $f(s)g(x) \leq f(x)g(x) \leq f(t)g(x)$, so $f(s)\int_a^b g(x)dx \leq \int_a^b f(x)g(x)dx \leq f(t)\int_a^b g(x)dx$ by Exercise 6.4. If $\int_a^b g(x)dx = 0$ then the result follows for any $c \in [a, b]$. Otherwise, $f(s) \leq \frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx} \leq f(t)$ so by the Intermediate Value Theorem we can find a point c between s and t or equal to s or t so that $f(c) = \frac{\int_a^b f(x)g(x)dx}{\int_a^b g(x)dx}$, so $\int_a^b f(x)g(x)dx = f(c)\int_a^b g(x)dx$.

□

Definition 34

If f is integrable on $[a, b]$ then we define $\int_b^a f = -\int_a^b f$. We also define $\int_a^a f = 0$ for any function f defined at a. For an interval $I = [a, b]$ we also use the notation $\int_a^b f = \int_I f$.

Note that in the latter notation, we always assume that $\int_I f$ is an integral where the lower limit (the subscript) on the integral is the least element of I and the upper integral bound or limit (the superscript) is the greatest element of I. Thus, if $b < a$ and I is the set of points between a and b or equal to a or b then $\int_I f = \int_b^a f$.

Theorem 6.13. *(a) Let $f : [a, b] \to \mathbb{R}$ be integrable, and let $c \in (a, b)$. Then $\int_a^c f(x)dx + \int_c^b f(x)dx = \int_a^b f(x)dx$.*

(b) If f is integrable on an interval containing a, b and c then $\int_a^c f(x)dx + \int_c^b f(x)dx = \int_a^b f(x)dx$.

Proof. (a) Let $\epsilon > 0$. Then we can find a partition P of $[a, b]$ so that $U(f, P) - L(f, P) < \epsilon$, and setting $Q = P \cup \{c\}$ we have $U(f, Q) - L(f, Q) < \epsilon$. If we define $P_1 = Q \cap [a, c]$ and $P_2 = Q \cap [c, b]$ then note that $U(f, Q) - L(f, Q) = (U(f, P_1) - L(f, P_1)) + (U(f, P_2) - L(f, P_2)) < \epsilon$, so $U(f, P_1) - L(f, P_1) < \epsilon$ and $U(f, P_2) - L(f, P_2) < \epsilon$. Thus, $\int_a^c f(x)dx$ and $\int_c^b f(x)$ exist.

Since $\int_a^c f + \int_c^b f, \int_a^b f \in (L(f,Q), U(f,Q))$ it follows that $|\int_a^c f + \int_c^b f - \int_a^b f| < \epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $\int_a^c f + \int_c^b f = \int_a^b f$.

(b) If $a < b < c$ then $\int_a^b f + \int_b^c f = \int_a^c f$, so $\int_a^b f = \int_a^c f - \int_b^c f = \int_a^c f + \int_c^b f$ by definition. Similarly, if $c < a < b$ then $\int_c^a f + \int_a^b f = \int_c^b f$, so $\int_a^b f = \int_c^b f - \int_c^a f = \int_a^c f + \int_c^b f$.

In the case where $a = b$ we just have $\int_a^c f + \int_c^b f = \int_a^c f + \int_c^a f = \int_a^c f - \int_a^c f = 0 = \int_a^b f$. If $a = c$ then $\int_a^c f + \int_c^b f = \int_a^a f + \int_a^b f = 0 + \int_a^b f = \int_a^b f$. If $c = b$ then $\int_a^c f + \int_c^b f = \int_a^b f + \int_b^b f = \int_a^b f + 0 = \int_a^b f$.

□

Theorem 6.14. *Let $f : [a,b] \to \mathbb{R}$ be integrable. Let $c \in [a,b]$ and let $F(x) = \int_c^x f(t)dt$. Then F is (uniformly) continuous.*

Proof. Choose M so that $|f(x)| < M$ for all $x \in [a,b]$. Let $x \in [a,b]$ and let $\epsilon > 0$. Choose $\delta = \frac{\epsilon}{M}$. If $|x - y| < \delta$ and $x, y \in [a,b]$ with $y < x$ then $|F(y) - F(x)| = |\int_x^y f|$ by Theorem 6.13. However $-M \leq f(x) \leq M$ for all $x \in [a,b]$ so $\int_x^y -M \leq \int_x^y f \leq \int_x^y M$ by Exercise 6.4, which means that $-M\delta \leq F(y) - F(x) < M\delta$. Hence, $|F(x) - F(y)| < M\frac{\epsilon}{M} = \epsilon$. Thus, F is uniformly continuous.

□

Theorem 6.15. *Fundamental Theorem of Calculus (first form). Let $f : [a,b] \to \mathbb{R}$ be continuous, let $c \in [a,b]$ and let $F(x) = \int_c^x f(x)dx$ for all $x \in [a,b]$. Then $F'(x) = f(x)$ if $a < x < b$. The function F is also continuous at a and b.*

Proof. First, continuity follows from Theorem 6.14. If $a = b$ then the result is vacuously true. Assume $a < b$ and let $x \in (a,b)$. If h has small enough absolute value then $x + h \in (a,b)$, and $\frac{F(x+h) - F(x)}{h} = \frac{\int_x^{x+h} f(x)dx}{h}$ by Theorem 6.13. By the First Mean Value Theorem for Integrals, there is some c_h between x and $x + h$ so that $\int_x^{x+h} f(x)dx = f(c_h)\int_x^{x+h} 1dx = hf(c_h)$. Hence, $\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{\int_x^{x+h} f(x)dx}{h} = \lim_{h\to 0} \frac{hf(c_h)}{h} = \lim_{h\to 0} f(c_h) = f(x)$ since f is continuous and $\lim_{h\to 0} c_h = x$ by the Squeeze Theorem. Hence, $F'(x) = f(x)$.

□

Theorem 6.16. *Fundamental Theorem of Calculus (second form). Let F be a function so that $F'(x) = f(x)$ for all $x \in [a,b]$, where $f(x)$ is integrable on $[a,b]$. Then $\int_a^b f(x)dx = F(b) - F(a)$.*

Proof. Let $\epsilon > 0$ and choose a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a,b]$ so that $U(f,P) - L(f,P) < \epsilon$. We know that F is differentiable (and thus continuous) on each interval $[x_{i-1}, x_i]$, so by the Mean Value Theorem, for each positive integer $i \leq n$ we can pick $x_i^* \in (x_{i-1}, x_i)$ so that $F'(x_i^*)(x_i - x_{i-1}) = F(x_i) - F(x_{i-1}) = f(x_i^*)(x_i - x_{i-1})$. Thus, with marking $T = \{x_1^*, ..., x_n^*\}$ we have $S_T(f,P) = \sum_{i=1}^n f(x_i^*)(x_i - x_{i-1}) = \sum_{i=1}^n F(x_i) - F(x_{i-1}) = F(b) - F(a)$. Since $\int_a^b f, F(b) - F(a) \in [L(f,P), U(f,P)]$, it must follow that $|\int_a^b f - (F(b) - F(a))| \leq U(f,P) - L(f,P) < \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $\int_a^b f = F(b) - F(a)$.

□

The following is essentially a slight generalization of the preceding theorem. The function F need not actually be differentiable at a and b as long as it is continuous at those points, and we could have $b < a$ or even $b = a$ and the theorem above would still be true. The proof is essentially the same as that of the Fundamental Theorem of Calculus second form, with minor adjustments.

Theorem 6.17. *Let F be a function so that $F'(x) = f(x)$ for all x between real numbers a and b, so that F is continuous at a and b, where $f(x)$ is integrable on the interval consisting of points a and b and all points in between those points. Then $\int_a^b f(x)dx = F(b) - F(a)$.*

Proof. First, assume $a < b$. Let $\epsilon > 0$ and choose a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a,b]$ so that $U(f,P) - L(f,P) < \epsilon$. We know that F is differentiable on each interval (x_{i-1}, x_i) and continuous on each interval $[x_{i-1}, x_i]$, so by the Mean Value Theorem, for each positive integer $i \leq n$ we can pick $x_i^* \in (x_{i-1}, x_i)$ so that $F'(x_i^*)(x_i - x_{i-1}) = F(x_i) - F(x_{i-1}) = f(x_i^*)(x_i - x_{i-1})$. Thus, with marking $T = \{x_1^*, ..., x_n^*\}$ we have $S_T(f,P) = \sum_{i=1}^n f(x_i^*)(x_i - x_{i-1}) = \sum_{i=1}^n F(x_i) - F(x_{i-1}) = F(b) - F(a)$. Since $\int_a^b f, F(b) - F(a) \in [L(f,P), U(f,P)]$, it must follow that $|\int_a^b f - (F(b) - F(a))| \leq U(f,P) - L(f,P) < \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $\int_a^b f = F(b) - F(a)$.

If $b < a$ then we know that $\int_b^a f = F(a) - F(b)$, so by definition $\int_a^b f = -\int_b^a f = F(b) - F(a)$. Likewise, if $a = b$ then $\int_a^a f = 0 = F(a) - F(a)$ so the result is true for all potential orderings of a and b.

□

We will not refer to the preceding theorem when we use the second form of the Fundamental Theorem of Calculus, but will just refer to the Fundamental Theorem of Calculus itself. Note that in the statement of the second form of the Fundamental Theorem of Calculus, f is listed as being integrable, but we also often see this theorem stated with f being continuous rather than just integrable (and we might see F simply as being differentiable on all of $[a, b]$ instead of just on the interior). If we refer to f as being continuous, then the second form of the Fundamental Theorem of Calculus is a direct consequence of the first form of the Fundamental Theorem of Calculus, which helps us understand why both theorems are called the Fundamental Theorem of Calculus. Here is how an argument would go for that version of the theorem.

Theorem 6.18. *Fundamental Theorem of Calculus (second form, continuous case). If $F'(x) = f(x)$ for all $x \in [a, b]$, where $a < b$, and $f(x)$ is continuous on $[a, b]$, then $\int_a^b f(x)dx = F(b) - F(a)$.*

Proof. We know f is integrable by Theorem 6.8. By the Fundamental Theorem of Calculus, first form, if we set $G(x) = \int_a^x f(t)dt$ then $G'(x) = f(x)$ for all $x \in (a, b)$ and we also know that G is continuous at a and b by Theorem 6.14. It is also true that $G(b) - G(a) = \int_a^b f(x)dx$. Since $F'(x) = G'(x)$ on (a, b), and F and G are both continuous at a and b, we know that $G(x) = F(x) + k$ for some constant k by Theorem 5.20. Thus, $\int_a^b f(x)dx = G(b) - G(a) = (F(b) + k) - (F(a) + k) = F(b) - F(a)$. □

The following is the typical form of u-substitution that is used in most calculus classes.

Theorem 6.19. *Basic u substitution. Let f and g be continuously differentiable functions so that $[a, b] \subset dom(f \circ g)$. Then $\int_a^b f'(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f'(u)du = f(g(b)) - f(g(a))$.*

Proof. Since f, g are continuously differentiable, we know that $f'(g(x))g'(x)$ is continuous and thus integrable. Since g is continuous, by the Intermediate Value Theorem, we know that all points between $g(a)$ and $g(b)$ are in $g([a, b])$ and hence in the domain of f. By the second form of the Fundamental Theorem of Calculus we know that $\int_{g(a)}^{g(b)} f'(u)du = f(g(b)) - f(g(a))$. Likewise, by the chain rule we know that $(f \circ g)'(x) = f'(g(x))g'(x)$, so by the Fundamental Theorem of Calculus we know that $\int_a^b f'(g(x))g'(x)dx = \int_a^b (f \circ g)'(x)dx = f(g(b)) - f(g(a)) = \int_{g(a)}^{g(b)} f'(u)du.$ □

Theorem 6.20. *Integration by Parts. If f and g are continuously differentiable on $[a,b]$ then $\int fg' = fg - \int gf'$ and $\int_a^b fg' = f(b)g(b) - f(a)g(a) - \int_a^b f'g$.*

Proof. First, note that by the product rule $(fg)' = f'g + g'f$ so $\int_a^b (fg)' = \int_a^b f'g + g'f$, so $\int fg' = fg - \int gf'$. Using the second form of the Fundamental Theorem of Calculus and Theorem 6.9, we have $f(b)g(b) - f(a)g(a) = \int_a^b f'g + g'f$, so $\int_a^b fg' = f(b)g(b) - f(a)g(a) - \int_a^b f'g.$

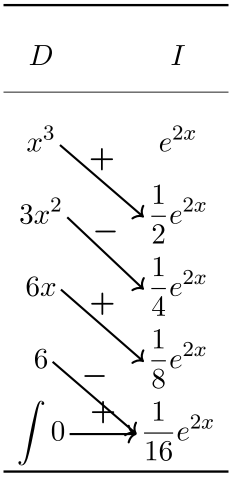
□

It is perhaps worth mentioning that keeping track of iterated uses of integration by parts is easier by way of using a table. You put the factor to be differentiated on the left, and the factor to be integrated on the right, and list iterated derivatives below the factor to be differentiated and iterated antiderivatives below the factor to be integrated. You then connect the left column entries with diagonal lines moving one row down to the right column entries representing multiplication, and put a plus or minus sign above the connecting line segments, starting with plus and alternating. Then when you reach a point in the table where you could multiply horizontally and get a product which is integrable, you make one final sign alteration on the last connecting (horizontal) line representing that product. You then write the sum of the indicated signs times the products indicated by the diagonal lines and add (or subtract depending on the sign that has been assigned) the final integral, which lets you leave the table and finish the problem with an integral.

Example 6.2. *Evaluate $\int x^3 e^{2x} dx$.*

Solution. Using the usual table (below) with x^3 to be differentiated until it becomes zero in the left column and e^{2x} to be integrated in the right column we end up integrating $(0)(e^{2x})$ to just get the constant of integration at the end, so the indefinite integral is $\int x^3 e^{2x} dx = \frac{1}{2}x^3 e^{2x} - \frac{3}{4}x^2 e^{2x} + \frac{3}{4}xe^{2x} - \frac{3}{8}e^{2x} + C.$

□



D		I
x^3	$+$	e^{2x}
$3x^2$	$-$	$\frac{1}{2}e^{2x}$
$6x$	$+$	$\frac{1}{4}e^{2x}$
6	$-$	$\frac{1}{8}e^{2x}$
$\int 0$	$+$	$\frac{1}{16}e^{2x}$

Taylor's series are one of the most helpful ways to analyze a function. This is based on integration by parts and induction as described below.

Theorem 6.21. *Taylor's Theorem (first form). Let n be a non-negative integer, and let $f^{(i)}(x)$ be continuous on the closed interval whose end points are x and a for natural numbers $i \le n+1$. Then $f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt$.*

Proof. We proceed by induction on n. For the $n = 0$ case, we note that $\int_a^x f'(t)dt = f(x) - f(a)$ by the Fundamental Theorem of Calculus, so it follows that $f(x) = \sum_{i=0}^{0} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{0!}\int_a^x f^{(0+1)}(t)(x-t)^0 dt$. Assume the result is true for $n = k$, so $f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{k!}\int_a^x f^{(k+1)}(t)(x-t)^k dt$. Then, using integration by parts, $\int_a^x f^{(k+1)}(t)(x-t)^k dt = \left.\frac{-f^{(k+1)}(t)(x-t)^{k+1}}{k+1}\right|_a^x + \frac{1}{k+1}\int_a^x f^{(k+2)}(t)(x-t)^{k+1}dt$, so $f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{k!}[\frac{f^{(k+1)}(a)(x-a)^{k+1}}{k+1} + \frac{1}{k+1}\int_a^x f^{(k+2)}(t)(x-t)^{k+1}dt] = \sum_{i=0}^{k+1} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{k+1!}\int_a^x f^{(k+2)}(t)(x-t)^{k+1}dt$ as desired. Thus, the result follows for all natural numbers n.

□

Theorem 6.22. *Taylor's Theorem (second form, Lagrange's error bound). Let n be a non-negative integer, and let $f^{(i)}(x)$ be continuous on the closed interval I whose end points are x*

and a for natural numbers $i \leq n+1$. Then $f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$ for some point $c \in I$.

Proof. Using the first form of Taylor's Theorem we know $f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(a)(x-a)^i}{i!} + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt$. Using the First Mean Value Theorem for integrals, since $f^{(n+1)}$ is continuous on I we know that for some point $c \in I$ it follows that $\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt = \frac{1}{n!} f^{(n+1)}(c) \int_a^x (x-t)^n dt = \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$.

□

Definition 35

We say that a function f is C^n on an open set U if $f^i(x)$ is continuous for all positive integers $i \leq n$. We say that f is C^∞ on U if $f^i(x)$ is continuous for all positive integers i. If U is the domain of f then we simply refer to f as C^n or C^∞ instead of C^n on U or C^∞ on U.

Definition 36

If $\{x_n\}$ is a sequence then the nth *partial sum* of this sequence is $s_n = \sum_{i=1}^{n} x_i$. The sequence $\{s_n\}$ is the *series* consisting of the sequence of partial sums s_n and is also denoted as $\sum_{n=1}^{\infty} x_n$. We also use $\sum_{n=1}^{\infty} x_n$ to refer to the point to which this series converges, depending on context.

Theorem 6.23. *Let $f : (c,d) \to \mathbb{R}$ be a C^∞ function so that $\{\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt\} \to 0$ for some a and x in (c,d). Then $f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(a)(x-a)^i}{i!}$. Furthermore, if $k_n \geq |f^{n+1}(x)|$ for all x between a and x and $\{\frac{k_n(x-a)^{n+1}}{(n+1)!}\} \to 0$ then $f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(a)(x-a)^i}{i!}$.*

Proof. By Theorem 6.21, we know that $\sum_{i=0}^{n}\frac{f^{(i)}(a)(x-a)^i}{i!}+\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt - f(x) = 0$, which means that if $\{\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt\} \to 0$ then it follows that $\{\sum_{i=0}^{n}\frac{f^{(i)}(a)(x-a)^i}{i!} - f(x)\} \to 0$ by Exercise 3.8, which means that $\{\sum_{i=0}^{n}\frac{f^{(i)}(a)(x-a)^i}{i!}\} \to f(x)$.

By Taylor's Theorem (second form), for some c between x and a or equal to x or a we know that $\frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!} = \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt$. Since $k_n \geq |f^{n+1}(x)|$ for all x between a and x (and thus at a and x as well since f^{n+1} is continuous), it follows that $|\frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}| \leq \frac{|k_n(x-a)^{n+1}}{(n+1)!}|$. Since $\{\frac{k_n(x-a)^{n+1}}{(n+1)!}\} \to 0$, it follows that $\frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt \to 0$ by the Squeeze Theorem, so by the argument above, $f(x) = \sum_{i=0}^{\infty}\frac{f^{(i)}(a)(x-a)^i}{i!}$. □

One of the many things we can illustrate with Taylor's series is the process called Newton's Method.

Newton's method a means of quickly approximating solutions to equations with a high level of accuracy. It is based on the notion that near a zero of a function, its tangent line will have an x-intercept close to the zero (since the curve and the tangent line should both be moving in approximately the same direction near a given point). Since it is easy to find the zero of the tangent line function, we can use this as our new starting point, take a tangent line at the point on the curve with x value equal to the zero of the tangent line, determine where the new tangent line to the curve intersects the x-axis, and repeat the process, becoming closer and closer to the zero.

To derive the formula for Newton's method, we begin with a guess at an initial approximation x_1 to a zero of a function $f(x)$. We then take the tangent line to the curve $y = f(x)$ at $(x_1, f(x_1))$. The slope of the tangent line is $f'(x_1)$, so the tangent line is $y - f(x_1) = f'(x_1)(x - x_1)$. Setting $f'(x_1)(x - x_1) + f(x_1) = 0$ we get $x = x_1 - \frac{f(x_1)}{f'(x_1)}$. Repeating this process, we motivate the following definition.

Definition 37

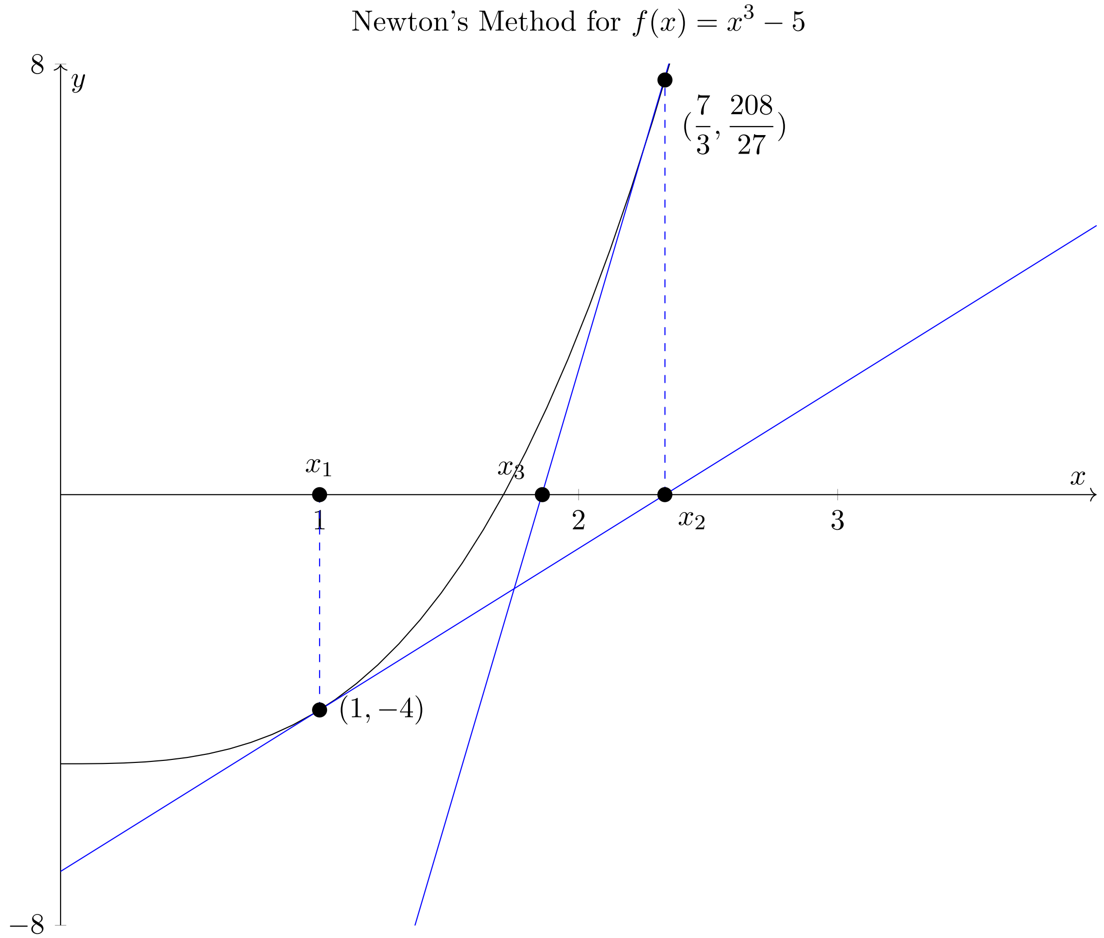
Inductively, if x_n is the nth Newton's method approximation to a zero of a differentiable function $f(x)$ containing x_n in its domain, then we define the $n+1$st approximation to be $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.

Note that in this definition, we did not define the initial guess x_1. That is a guess which can be made based on something like the Intermediate Value Theorem, just a stab in the dark at where a zero might be, or by looking at the graph and trying to eyeball around where a zero might be. For purposes of this process, we do not have an established algorithm for the primary guess. Below is a picture of showing how Newton's method works for the specific example of trying to estimate a zero of $f(x) = x^3 - 5$. This is not interesting in terms of showing Newton's method's strength for estimating roots quickly since we know that the only zero of the function is $5\frac{1}{3}$. However, since we know the solution, this does help us to compare the estimates to the approximations.

Example 6.3. *Starting with an initial approximation of $x_1 = 1$ to the zero of $f(x) = x^3 - 5$, find x_2 and x_3 using Newton's method and approximate the cube root of five. You may use decimal approximations to your answers.*

Solution. We begin with an initial guess, $x_1 = 1$. This estimate was picked solely on the basis that we like the number 1 (pretty much a wild guess). Sure enough, 1 is not a zero of the function. We take $f'(x) = 3x^2$. Plugging into the formula for Newton's method we get $x_2 = 1 - \frac{1^3 - 5}{3(1^2)} = \frac{7}{3}$. To get the next estimate we take $x_3 = \frac{7}{3} - \frac{\frac{343}{27} - \frac{135}{27}}{\frac{49}{3}}$. A decimal approximation to this estimate is 1.86. This is still fairly far from the true value of the cube root, which is about 1.709975947. However, if we do one more iteration we will get about 1.72. So, have about four iterations we have become quite close to the true value of the root. One more iteration gives us 1.710065. This is still off, of course, but it is only off by about one ten thousandth, so it is pretty close. Doing one more iteration gives us 1.709975951.

□



Notice that the decimals all agree until the eighth place past the decimal. Here is a picture for the first two steps of the estimate. The initial estimate was pretty far off, and the derivative was fairly small, so the second estimate wasn't great, but by the third estimate the estimates were starting to be pretty close to the cube root of five. Newton's method works better when the derivative is not close to zero at the approximation point, particularly if the associated function value at the approximation point is large. If the derivative actually is zero at an approximation then you can't use Newton's method at that point (you would be dividing by zero). There are also anomalies where functions can actually bounce back and forth between two estimates indefinitely using Newton's method and never get you any closer to the actual zero. Also, if you pick a point that isn't close to the zero you are trying to find, you may wall get a sequence of approximations that converges to a zero of the function which is a different zero from the zero you were hoping to approximate (assuming the function has multiple zeros). However, most of the time this method will estimate a zero quite accurately and quite quickly, making it much better than the other methods we have discussed up to this point for estimating zeroes of functions. If we pick an initial estimate near the zero we want, then Newton's method will generally give approximations converging to that particular zero. What we want, more specifically, is for the second derivative to be reasonably small and the first derivative to be reasonably large, and to start out at an approximating value which is near the zero we hope to find, in which case we approach the zero rapidly. Here is a more specific error bound.

Theorem 6.24. *Error bound for Newton's method. Let f, f', f'' be continuous on an interval I containing x_n, x_{n+1} and r, where x_n and x_{n+1} are the nth and $n{+}1$st approximations using Newton's method for the zero r of the function $f(x)$. If $L, M > 0$ and for all $x \in I$ it is true that $|f'(x)| \geq L$ and $|f''(x)| \leq M$, then it follows that $|x_{n+1} - r| \leq \frac{M}{2L}|x_n - r|^2$. If $\frac{M}{2L}|x_1 - r| \leq \frac{1}{2}$ then $|x_n - r| < \frac{2L}{M}(\frac{1}{2})^{2^{n-1}}$.*

Proof. By Taylor's Theorem with the second (Lagrange's) remainder formula we know that $f(x) = f(x_n) + f'(x_n)(x - x_n) + R$ on the interior of I, where $R = \frac{f''(c)(x-x_n)^2}{2!}$ for some c between x and x_n. Hence $|R| \leq \frac{M}{2}|x - x_n|^2$. If $x = r$ then we get $0 = f(r) = f(x_n) + f'(x_n)(r - x_n) + R$, so $r - x_n + \frac{f(x_n)}{f'(x_n)} = -\frac{R}{f'(x_n)}$. Since $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ this gives us that $|x_{n+1} - r| = \frac{|R|}{|f'(x_n)|} \leq \frac{M}{2L}|x_n - r|^2$.

In the case where $\frac{M}{2L}|x_1 - r| \leq \frac{1}{2}$, this means that $|x_2 - r| \leq \frac{M}{2L}|x_1 - r|^2 \leq (\frac{1}{2})(|x_1 - r|) \leq \frac{L}{M}(\frac{1}{2}) < \frac{2L}{M}(\frac{1}{2})$. Proceeding by induction, we assume that $|x_k - r| < (\frac{2L}{M})(\frac{1}{2})^{2^{k-1}}$. Then we have $|x_{k+1} - r| \leq \frac{M}{2L}|x_k - r|^2 < \frac{M}{2L}(\frac{2L}{M})^2(\frac{1}{2})^{2^{k+1-1}} = \frac{2L}{M}(\frac{1}{2})^{2^{k+1-1}}$. The result follows by induction.

□

The first form of the Fundamental Theorem of Calculus makes certain function derivatives more accessible than they would be without integration.

Definition 38

Define $\ln(x) = \int_1^x \frac{1}{t}dt$ for all $x > 0$.

Note that $\ln(x)$ exists by Theorem 6.8.

Theorem 6.25. *For all $x, y > 0$ and $r \in \mathbb{Q}$ the following are true:*

(a) $\ln(x)' = \frac{1}{x}$ *for each* $x > 0$.
(b) $\ln(x)$ *is increasing and* $\ln(1) = 0$.
(c) $\ln(xy) = \ln(x) + \ln(y)$
(d) $\ln(\frac{x}{y}) = \ln(x) - \ln(y)$
(e) $\ln(x^r) = r\ln(x)$ *if* $r \in \mathbb{Q}$
(f) $\ln(x)$ *has no lower bound or upper bound. The range of* $\ln(x)$ *is* $\mathbb{R}$.
(g) There is a unique number which we will refer to as e so that $\ln(e) = 1$

(h) We define $\exp(x)$ *to be the inverse of* $\ln(x)$*. Then* $\exp(x)$ *is increasing on* $\mathbb{R}$ *and* $(\exp(x))' = \exp(x)$*.*

(i) If $r = \dfrac{p}{q}$*, where* p *is an integer and* q *is a natural number then* $e^r = \exp(r)$*.*

Proof. (a) This follows from the Fundamental Theorem of Calculus (version 1).

(b) $\ln(x)$ is increasing by Theorem 5.18 since $(\ln(x))' = \dfrac{1}{x} > 0$, and $\ln(1) = \displaystyle\int_1^1 \frac{1}{t}dt = 0$ by definition.

(c) Note that (treating y as constant), $(\ln(xy))' = \dfrac{y}{xy} = \dfrac{1}{x} = (\ln(x))'$. Thus, by Theorem 5.20, $\ln(x) + k = \ln(xy)$ for some constant k. Setting $x = 1$ we have $k = \ln(y)$.

(d) By (c) we know that $0 = \ln(1) = \ln(y\frac{1}{y}) = \ln(y) + \ln(\frac{1}{y})$, so $\ln(\frac{1}{y}) = -\ln(y)$. Hence, $\ln(\frac{x}{y}) = \ln(x) + \ln(\frac{1}{y}) = \ln(x) - \ln(y)$.

(e) $r = \dfrac{p}{q}$ for some integers p and q, with $q > 0$. Inductively, we note that $\ln(x^1) = \ln(x)$ and if $\ln(x^k) = k\ln(x)$ then $\ln(x^{k+1}) = \ln(x^k x) = \ln(x^k) + \ln(x) = k\ln(x) + \ln(x) = (k+1)\ln(x)$, so for any natural number n it follows that $\ln(x^n) = n\ln(x)$. Likewise, $\ln(x) = \ln((x^{\frac{1}{n}})^n) = n\ln(x^{\frac{1}{n}})$, so $\ln(x^{\frac{1}{n}}) = \dfrac{1}{n}\ln(x)$ for any natural number n. If m is a negative integer then $\ln(x^m) = \ln(\dfrac{1}{x^{-m}}) = -(-m\ln(x)) = m\ln(x)$. If $r = 0$ then $\ln(x^r) = \ln(1) = 0 = r\ln(x)$.

Combining these, we have that $\ln(x^{\frac{p}{q}}) = p\ln(x^{\frac{1}{q}}) = \dfrac{p}{q}\ln(x)$.

(f) By (b) we know $\ln(2) > \ln(1) = 0$. Since the natural numbers are not bounded above, for any $M > 0$ we can find a natural number k so that $k > \dfrac{M}{\ln(2)}$, so $k\ln(2) > M$, and thus $\ln(2^k) > M$, which means $\ln(x)$ is not bounded above. Similarly, $\ln(2^{-k}) = -k\ln(2) < -M$, so $\ln(x)$ is not bounded below.

For any $z \in \mathbb{R}$ there are, thus, $a, b > 0$ so that $\ln(a) < z < \ln(b)$ and hence, by the Intermediate Value Theorem, there is some point c between a and b so that $\ln(c) = b$, so every real number is in the range of $\ln(x)$.

(g) By (f) we can find a point e so that $\ln(e) = 1$. This is the only number whose natural logarithm is one because $\ln(x)$ is increasing by (b) and therefore one to one.

(h) We know that $\exp(x)$ has a domain of all real numbers because $\mathbb{R}$ is the range of $\ln(x)$ by (f). By the Inverse Function Theorem, $\exp(x)$ is increasing and differentiable. Hence, setting $y = \exp(x)$, we know $\ln(y) = x$ where y is differentiable, so by the chain rule $\dfrac{1}{y}y' = 1$, which means that $y = y'$ and $\exp(x) = (\exp(x))'$.

(i) Note that $\ln(e^r) = r\ln(e) = r$ by (e), and $\ln(\exp(r)) = r$ by definition of inverse. Since $\ln(x)$ is one to one, it follows that $e^r = \exp(r)$.

□

Definition 39

Let $x > 0$ and let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Then we define $x^\alpha = \exp(\alpha \ln(x))$.

Theorem 6.26. *(a)* $e^x = \exp(x)$ *for all* $x \in \mathbb{R}$

(b) For all $x \in (0, \infty)$, $(x^r)' = rx^{r-1}$

(c) For each $r > 0$, $(r^x)' = \ln(r)r^x$.

Proof. (a) By definition, $e^x = \exp(x\ln(e)) = \exp(x)$

(b) $(x^r)' = (e^{r\ln(x)})' = \dfrac{r}{x}e^{r\ln(x)} = \dfrac{r}{x}x^r = rx^{r-1}$ by the chain rule (and Exercise 6.12).

(c) Using the chain rule, we know that $(r^x)' = \exp(x\ln(r))' = \ln(r)\exp(x\ln(r)) = \ln(r)r^x$.

□

The Fundamental Theorem of Calculus shows us how an antiderivative can be used to evaluate an integral, but there are some differences between the idea of an antiderivative, the most general antiderivative, a definite integral and an indefinite integral.

Definition 40

We say that $g(x)$ is an *antiderivative* of $f(x)$ on (a, b) if $g'(x) = f(x)$ for all x in (a, b).

Note that, by Theorem 5.20, if $g(x)$ and $h(x)$ are antiderivatives of a function $f(x)$ on an interval I then $g(x) = h(x) + c$ (for some constant c) on the interval I.

Definition 41

We refer to a collection of functions $\mathcal{C}$ as the *most general antiderivative* of a continuous function $f(x)$ if every antiderivative of f is an element of $\mathcal{C}$. The notation $\displaystyle\int f(x)dx = g(x)+C$ means that the set of all functions $g(x)+C$ so that $C \in \mathbb{R}$ is the *indefinite integral* of $f(x)$. This means that, on every open interval $(a, b) \subseteq dom(f)$, the most general antiderivative of the function f restricted to domain (a, b) is the set of all functions of the form $g(x) + C$ so that $C \in \mathbb{R}$.

Note that the indefinite integral is not necessarily the most general antiderivative of f, only the form of the most general antiderivative of f on each open interval on which f is continuous.

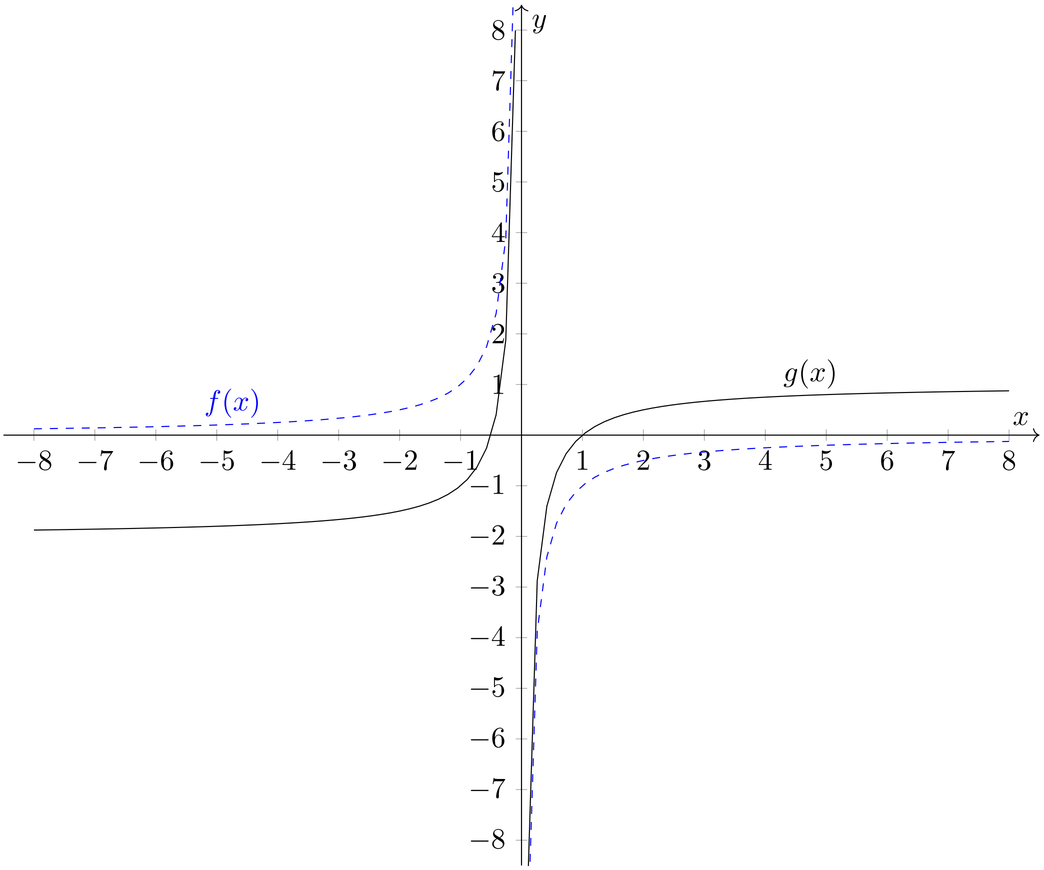
For functions which are continuous on a domain consisting of a finite set of mutually exclusive open intervals, the most general antiderivative is obtained by taking an antiderivative over each such interval on which f is defined and adding different constants. This is illustrated in the following example:

Example 6.4. *Find the most general antiderivative of* $f(x) = \dfrac{1}{x^2}$.

Solution: First, we note that the derivative of $\dfrac{-1}{x}$ is $\dfrac{1}{x^2}$. Since $f(x)$ is continuous on $(0,\infty)$ and on $(-\infty,0)$ we conclude that $F(x) = -\dfrac{1}{x} + C_1$ when $x > 0$ and $F(x) = -\dfrac{1}{x} + C_2$ when $x < 0$ is the most general antiderivative of $f(x)$ (where the constants need not be equal). In other words, every antiderivative of $f(x)$ is of the form stated for $F(x)$.

This is a picture of the graph of an antiderivative of $\dfrac{1}{x^2}$. The function $\dfrac{-1}{x} + C$ is the indefinite integral, so we could add different constants on each of the two components of the domain of $\dfrac{1}{x^2}$. The particular antiderivative we will graph is $g(x) = -2 - \dfrac{1}{x}$ if $x < 0$ and $g(x) = 1 - \dfrac{1}{x}$ if $x > 0$. The blue dashed graphs represent the function $f(x) = -\dfrac{1}{x}$ for comparison. Notice that the slopes at every point are the same as the slope of the graph of $y = \dfrac{1}{x}$, so the derivative is the same as the derivative of $y = \dfrac{1}{x}$.

Antiderivative of $f(x) = \dfrac{1}{x^2}$



Using the sum rule for derivatives and working backwards we get that if $g_1, g_2, .., g_n$ are continuous functions with antiderivatives $f_1, f_1, ..., f_n$ on each open interval in the domain of $G(x) = \alpha_1 g_1(x) + \alpha_2 g_2(x) + ...\alpha_n g_n(x)$, then $\int \alpha_1 g_1(x) + \alpha_2 g_2(x) + ...\alpha_n g_n(x)dx = \alpha_1 f_1(x) + \alpha_2 f_2(x) + ...\alpha_n f_n(x) + C$. For example, $\int x^3 + 4x + 9dx = \frac{x^4}{4} + 2x^2 + 9x + C$.

Theorem 6.27. *Let $f(x) = \frac{1}{x}$. The most general antiderivative of $\frac{1}{x}$ is the set of all functions $F(x)$ of the form $F(x) = \ln|x| + C_1$ if $x > 0$ and $F(x) = \ln|x| + C_2$ if $x < 0$, where $C_1, C_2 \in \mathbb{R}$.*

The indefinite integral $\int \frac{1}{x} = \ln|x| + C$.

Proof. If $x > 0$ then $\ln(x)$ is defined and we know that $(\ln(x))' = \frac{1}{x}$, which means that every antiderivative of $f(x)$ restricted to any open interval $(a, b) \subset (0, \infty)$ has the form $\ln(x) + C$ for some constant C by Theorem 5.20. If $x < 0$, however, then $\frac{1}{x}$ still exists, and $\ln(-x)$ is defined. Using the chain rule we see that $(\ln(-x))' = \frac{1}{-x}(-1) = \frac{1}{x}$. As before, for $(a, b) \subset (-\infty, 0)$ each antiderivative of $f(x)$ restricted to (a, b) has the form $\ln(-x) + C = \ln|x| + C$. we have $\int \frac{1}{x}dx = \ln(-x) + C_2$. Since $-x = |x|$ if $x < 0$ we can consolidate this notation by saying that the most general antiderivative of $\frac{1}{x}$ is $\ln|x| + C_1$ if $x > 0$ and $\ln|x| + C_2$ if $x < 0$. □

Note that in the preceding example we demonstrated that $\int \frac{1}{x}dx = \ln|x| + C$ is a correct formula for the indefinite integral (whether we consider the indefinite integral to be the general antiderivative on an interval wherein $x > 0$ or an interval where $x < 0$).

The difference between ”most general antiderivative” and ”indefinite integral” is also not always immediately clear. So, $\int \frac{1}{x}dx = \ln|x| + C$ means that if we took an open interval (a, b) contained in the domain of $f(x) = \frac{1}{x}$, which would have to be a subset of either $(-\infty, 0)$ or $(0, \infty)$ since the natural log of zero is undefined, then the set of all functions of the form $\ln|x| + C$ for $C \in \mathbb{R}$ would be the most general antiderivative for f if the domain of f were restricted to (a, b). This does not mean that the most general antiderivative of f on its entire domain is the same set of functions, only that if we restricted f to an open interval contained in its domain then all antiderivatives of f on that open interval would be of that form. If $(a, b) \subseteq (0, \infty)$, for instance, then all antiderivatives of $f(x)$ on (a, b) would have form $\ln(x) + C = \ln|x| + C$. If $(a, b) \subseteq (-\infty, 0)$ then all antiderivatives of f would have form $\ln(-x) + C = \ln|x| + C$ for some constant C. However, an antiderivative of f on its entire domain (not an open interval subset of it) could have form $\ln|x| + C_1$ if $x > 0$ and $\ln|x| + C_2$ if $x < 0$, where C_1 and C_2 are different constants (so the indefinite integral does not include all antiderivatives of the function in this case). When the domain of f is a connected interval then the indefinite integral and the most general antiderivative of f are the same thing. When the domain is a union of disconnected intervals then a different constant could be added to the antiderivative on each component of the domain and still result in an antiderivative for the function, which means that that most general antiderivative of $f(x)$ and the indefinite integral are not the same thing in that case.

Exercises:

Exercise 6.1. *Let $f : [a,b] \to \mathbb{R}$ be integrable and let $F(x) = \int_a^x f(t)dt$. Then F is integrable.*

Exercise 6.2. *Let $f(x) = 0$ if $x \neq 0$ and $f(x) = 1$ if $x = 0$. Prove that $\int_{-1}^{1} f(x)dx = 0$.*

Exercise 6.3. *If $f : [a,b] \to \mathbb{R}$ is bounded and has only finitely many discontinuities then f is integrable.*

Exercise 6.4. *(a) Let f, g be integrable on $[a,b]$ and let $f(x) \leq g(x)$ for all $x \in [a,b]$. Then $\int_a^b f(x)dx \leq \int_a^b g(x)dx$.*

(b) Let f be integrable on $[a,b]$. If $m \leq f(x) \leq M$ for all $x \in [a,b]$ then $m(b-a) = \int_a^b mdx \leq \int_a^b f(x)dx \leq \int_a^b Mdx = M(b-a)$.

Exercise 6.5. *Let $f : [a,b] \to \mathbb{R}$ be continuous and non-negative. Prove that if $f(c) > 0$ for some $c \in [a,b]$ then $\int_a^b f(x)dx > 0$.*

Exercise 6.6. *Let $f : [a,b] \to \mathbb{R}$ be monotone. Then f is integrable.*

Exercise 6.7. *Let $g_1(x), g_2(x)$ be differentiable on $[a,b]$, let $f(x)$ be continuous on $[a,b]$ and let $F(x) = \int_{g_1(x)}^{g_2(x)} f(t)dt$. Then $F'(x) = f(g_2(x))g_2'(x) - f(g_1(x))g_1'(x)$.*

Exercise 6.8. *Show that $\lim_{n\to\infty} (1 + \frac{1}{n})^n = e$.*

Exercise 6.9. *Let $f, g : [a,b] \to \mathbb{R}$ be integrable functions. Then fg is bounded.*

Exercise 6.10. *Define $\frac{e^x + e^{-x}}{2} = \cosh(x)$ and $\frac{e^x - e^{-x}}{2} = \sinh(x)$. Then $(\sinh(x))' = \cosh(x)$ and $(\cosh(x))' = \sinh(x)$.*

Exercise 6.11. *Find, with proof, an example of a function $f(x)$ which is integrable on $[a, b]$ so that $F(x) = \int_a^x f(t)dt$ is not differentiable.*

Exercise 6.12. *Prove that if $c > 0$ and $a, b \in \mathbb{R}$ then:*

(a) $c^a c^b = c^{a+b}$.

(b) $\frac{c^a}{c^b} = c^{a-b}$

(c) $(c^a)^b = c^{ab}$.

Exercise 6.13. *Prove that if $\{x_n\}$ is a sequence of positive numbers converging to a real number r and $c > 0$ then $\{c^{x_n}\} \to c^r$.*

Note that, as a result of this exercise, if we were to define exponents at irrational numbers as the limits of exponents at the first n digits of the decimal expansions of those numbers, then the definition of raising a number to an irrational number would be equivalent to the definition given above.

Exercise 6.14. *Let $f : [a, b] \to \mathbb{R}$ be integrable. Then $|f|$ is integrable, and $|\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx$.*

Hints:

Hint to Exercise 6.1. *Let $f : [a, b] \to \mathbb{R}$ be integrable and let $F(x) = \int_a^x f(t)dt$. Then F is integrable.*

Continuous functions are integrable.

Hint to Exercise 6.2. *Let $f(x) = 0$ if $x \neq 0$ and $f(x) = 1$ if $x = 0$. Prove that $\int_{-1}^1 f(x)dx = 0$.*

For a given $\epsilon > 0$, find a partition so that the upper sum over the partition is ϵ and the lower sum is zero.

Hint to Exercise 6.3. *If $f : [a, b] \to \mathbb{R}$ is bounded and has only finitely many discontinuities then f is integrable.*

If the supplementary material was covered, use the Lebesgue Characterization of Riemann Integrability. Otherwise, find upper and lower sums within distance ϵ of each other by picking a partition with short subinterval rectangles about each point of discontinuity.

Hint to Exercise 6.4. *(a) Let f, g be integrable on $[a, b]$ and let $f(x) \leq g(x)$ for all $x \in [a, b]$. Then $\int_a^b f(x)dx \leq \int_a^b g(x)dx$.*

(b) Let f be integrable on $[a, b]$. If $m \leq f(x) \leq M$ for all $x \in [a, b]$ then $m(b - a) = \int_a^b mdx \leq \int_a^b f(x)dx \leq \int_a^b Mdx = M(b - a)$.

Use the Riemann sum characterization of integral (Theorem 6.7). Take a sequence of partitions with mesh converging to zero, use any markings you wish and then compare the Riemann sums for f and g using those markings.

Hint to Exercise 6.5. *Let $f : [a, b] \to \mathbb{R}$ be continuous and non-negative. Prove that if $f(c) > 0$ for some $c \in [a, b]$ then $\int_a^b f(x)dx > 0$.*

Since f is continuous, it is possible to guarantee that f is larger than some positive value on some open interval centered at c. Find a partition with a subinterval containing c which is sufficiently small.

Hint to Exercise 6.6. *Let $f : [a, b] \to \mathbb{R}$ be monotone. Then f is integrable.*

Recall that the supremum on a subinterval induced by a partition is always the function value at the right end point for an increasing function, and the infimum the value of the function at the left end point. What form does the upper minus lower sum take?

Hint to Exercise 6.7. *Let $g_1(x), g_2(x)$ be differentiable on $[a, b]$, let $f(x)$ be continuous on $[a, b]$ and let $F(x) = \int_{g_1(x)}^{g_2(x)} f(t)dt$. Then $F'(x) = f(g_2(x))g_2'(x) - f(g_1(x))g_1'(x)$.*

Combine the Fundamental Theorem of Calculus (first form) with the chain rule.

Hint to Exercise 6.8. *Show that $\lim_{n\to\infty}(1 + \frac{1}{n})^n = e$.*

You could also proceed by taking the log and applying Theorem 4.11, or you could use the definition of $\ln(x)$ and the definition of e directly.

Hint to Exercise 6.9. *Let $f, g : [a, b] \to \mathbb{R}$ be integrable functions. Then fg is bounded.*

Recall that integrable functions are bounded.

Hint to Exercise 6.10. *Define $\frac{e^x + e^{-x}}{2} = \cosh(x)$ and $\frac{e^x - e^{-x}}{2} = \sinh(x)$. Then $(\sinh(x))' = \cosh(x)$ and $(\cosh(x))' = \sinh(x)$.*

Use the chain rule.

Hint to Exercise 6.11. *Find, with proof, an example of a function $f(x)$ which is integrable on $[a, b]$ so that $F(x) = \int_a^x f(t)dt$ is not differentiable.*

The function$f$ would have to be discontinuous. Look at functions with a jump discontinuity.

Hint to Exercise 6.12. *Prove that if $c > 0$ and $a, b \in \mathbb{R}$ then:*
(a) $c^a c^b = c^{a+b}$.
(b) $\frac{c^a}{c^b} = c^{a-b}$
(c) $(c^a)^b = c^{ab}$.

Take the logs of both sides, and recall that $\ln(x)$ is one to one.

Hint to Exercise 6.13. *Prove that if $\{x_n\}$ is a sequence of positive numbers converging to a real number r and $c > 0$ then $\{c^{x_n}\} \to c^r$.*

Start by taking the natural log of the sequence terms and use Theorem 4.11.

Hint to Exercise 6.14. *Let $f : [a, b] \to \mathbb{R}$ be integrable. Then $|f|$ is integrable, and $|\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx$.*

Try using theorems 6.10, 1.16 and 6.4.

Solutions:

Solution to Exercise 6.1. *Let $f : [a,b] \to \mathbb{R}$ be integrable and let $F(x) = \displaystyle\int_a^x f(t)dt$. Then F is integrable.*

Proof. By Theorem 6.14, we know that F is continuous, which means that F is integrable by Theorem 6.8. □

Solution to Exercise 6.2. *Let $f(x) = 0$ if $x \neq 0$ and $f(x) = 1$ if $x = 0$. Prove that $\displaystyle\int_{-1}^{1} f(x)dx = 0$.*

Proof. Let $\epsilon > 0$ and let $P = \{-1, -\frac{\epsilon}{4}, \frac{\epsilon}{4}, 1\}$. Then $L(f,P) = 0$ and $U(f,P) = \frac{\epsilon}{2}$, so $U(f,P) - L(f,P) < \epsilon$ and f is integrable. Since $L(f,Q) = 0$ for every partition Q it must be the case that $(L)\displaystyle\int_a^b f = \int_a^b f = 0$. □

Solution to Exercise 6.3. *If $f : [a,b] \to \mathbb{R}$ is bounded and has only finitely many discontinuities then f is integrable.*

Proof. Since finite sets have Lebesgue measure zero by 7.61, f is integrable by Theorem 7.73. □

OR without the supplementary materials:

Proof. Let $\epsilon > 0$. Let $a \leq x_1 < x_2 < ... < x_m \leq b$, where $D = \{x_1, x_2, x_3, ..., x_m\}$ is the set of points at which f is discontinuous. Since f is bounded we can choose $M > 0$ so that $|f(x)| < M$ for all $x \in [a,b]$. Let $Q = \{q_0, q_1, q_2, ..., q_n\}$ be a partition of $[a,b]$ whose mesh is less than $\frac{\epsilon}{8Mm}$. Let $K = \displaystyle\bigcup_{\{i \in \{1,2,...,n\} | [q_{i-1}, q_i] \cap D = \emptyset\}} [q_{i-1}, q_i]$. Then K is contained in $[a,b]$ and is bounded, and K is the union of finitely many closed intervals so K is closed. Since f is continuous at every point of K, we know that f is uniformly continuous on K by Theorem 4.18, so we can find a $\delta > 0$ so that if $|x - y| < \delta$ and $x, y \in K$ then $|f(x) - f(y)| < \frac{\epsilon}{2(b-a)}$. Let $P = \{p_0, p_1, p_2, ..., p_k\}$ be a partition of $[a,b]$ which is a refinement of Q with mesh less than δ. At most two subintervals induced by P can contain any discontinuity x_i, which means that no more than $2m$ subintervals induced by P intersect D. $U(f,P) - L(f,P) = \displaystyle\sum_{\{i \in \{1,2,...,k\} | [p_{i-1}, p_i] \cap D = \emptyset\}} (M_i - m_i)(p_i - p_{i-1}) + \sum_{\{i \in \{1,2,...,k\} | [p_{i-1}, p_i] \cap D \neq \emptyset\}} (M_i - m_i)(p_i - p_{i-1})$. Since $\displaystyle\sum_{\{i \in \{1,2,...,k\} | [p_{i-1}, p_i] \cap D \neq \emptyset\}} (M_i - m_i)(p_i - p_{i-1}) < (2M)(2m)(\frac{\epsilon}{8Mm}) = \frac{\epsilon}{2}$, and $\displaystyle\sum_{\{i \in \{1,2,...,k\} | [p_{i-1}, p_i] \cap D = \emptyset\}} (M_i - m_i)(p_i - p_{i-1}) < (b-a)(\frac{\epsilon}{2(b-a)}) = \frac{\epsilon}{2}$, it follows that $U(f,P) - L(f,P) < \epsilon$, so f is integrable.

□

Solution to Exercise 6.4. *(a) Let f, g be integrable on $[a, b]$ and let $f(x) \le g(x)$ for all $x \in [a, b]$. Then $\int_a^b f(x)dx \le \int_a^b g(x)dx$.*

(b) Let f be integrable on $[a, b]$. If $m \le f(x) \le M$ for all $x \in [a, b]$ then $m(b - a) = \int_a^b mdx \le \int_a^b f(x)dx \le \int_a^b Mdx = M(b - a)$.

Proof. Let $\{P_n\}$ be a sequence of partitions of $[a, b]$ with $\{|P_n|\} \to 0$, and choose a marking T_n for each P_n. Then since $f(x) \le g(x)$ it follows that $S_{T_n}(f, P_n) \le S_{T_n}(g, P_n)$. Since f, g are integrable, we have proven that $\{S_{T_n}(f, P_n)\} \to \int_a^b f(x)dx$ and $\{S_{T_n}(g, P_n)\} \to \int_a^b g(x)dx$. By the Comparison Theorem, $\int_a^b f(x)dx \le \int_a^b g(x)dx$.

Let k be any real number and $P = \{x_0, x_1, ..., x_m\}$ be a partition of $[a, b]$. Then $L(f, P) = U(f, P) = \sum_{i=1}^{m} k(x_i - x_{i-1}) = k(b - a)$ which means that $(U)\int_a^b k = (L)\int_a^b k = \int_a^b = k(b-a)$. Hence, if $m \le f(x) \le M$ for all $x \in [a, b]$ then $m(b-a) = \int_a^b mdx \le \int_a^b f(x)dx \le \int_a^b Mdx = M(b - a)$. □

Solution to Exercise 6.5. *Let $f : [a, b] \to \mathbb{R}$ be continuous and non-negative. Prove that if $f(c) > 0$ for some $c \in [a, b]$ then $\int_a^b f(x)dx > 0$.*

Proof. Since f is continuous, f is integrable by Theorem 6.8, and we may choose $0 < \delta$ so that if $|x - c| < \delta$ then $|f(x) - f(0)| < \frac{f(c)}{2}$, so $f(x) > \frac{f(c)}{2}$. Then let P be a partition of $[a, b]$ with $|P| < \delta$. There is some subinterval $[x_{j-1}, x_j]$ induced by P which contains c. Thus, $\int_a^b f \ge L(f, P) \ge (x_j - x_{j-1})(\frac{f(c)}{2}) > 0$.

□

Solution to Exercise 6.6. *Let $f : [a, b] \to \mathbb{R}$ be monotone. Then f is integrable.*

Proof. Let $\epsilon > 0$. First, assume that f is non-decreasing. Choose a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$ so that $|P| < \frac{\epsilon}{f(b) - f(a) + 1}$. Since f is non-decreasing, the infimum and supremum of all $f(x)$ values on the subinterval $[x_{i-1}, x_i]$ are $f(x_{i-1})$ and $f(x_i)$ respectively, for each $1 \le i \le n$. Hence $U(f, P) - L(f, P) = \sum_{i=1}^{n}(M_i - m_i)(x_i - x_{i-1}) = \sum_{i=1}^{n}(f(x_i) - f(x_{i-1}))(x_i - x_{i-1}) < \frac{\epsilon}{f(b) - f(a) + 1}\sum_{i=1}^{n}(f(x_i) - f(x_{i-1})) = \frac{\epsilon}{f(b) - f(a) + 1}(f(b) - f(a)) < \epsilon$. Thus, f is integrable. □

Solution to Exercise 6.7. *Let $g_1(x), g_2(x)$ be differentiable on $[a,b]$, let $f(x)$ be continuous on $[a,b]$ and let $F(x) = \int_{g_1(x)}^{g_2(x)} f(t)dt$. Then $F'(x) = f(g_2(x))g_2'(x) - f(g_1(x))g_1'(x)$.*

Proof. Since $F(x) = \int_{g_1(x)}^{g_2(x)} f(t)dt = \int_a^{g_2(x)} f(t)dt - \int_a^{g_1(x)} f(t)dt$, it follows from the chain rule and the First Form of the Fundamental Theorem of Calculus that $F'(x) = f(g_2(x))g_2'(x) - f(g_1(x))g_1'(x)$.

□

Solution to Exercise 6.8. *Show that $\lim_{n\to\infty}(1+\frac{1}{n})^n = e$.*

Proof. We know that $\ln(1+\frac{1}{n})^n = n\ln(1+\frac{1}{n}) = n\int_1^{1+\frac{1}{n}} \frac{1}{t}dt$. Since $\frac{1}{t}$ is decreasing, the largest value of $\frac{1}{t}$ on $[1, 1+\frac{1}{n}]$ is 1 and the smallest value is $\frac{n}{n+1}$. Thus, by Exercise 6.4 we know that $\frac{1}{n}\frac{n}{n+1} \leq \int_1^{1+\frac{1}{n}} \frac{1}{t}dt \leq \frac{1}{n}(1)$, which means that $\frac{n}{n+1} \leq n\ln(1+\frac{1}{n}) \leq 1$. Hence, by the Squeeze Theorem we know that $\ln(1+\frac{1}{n})^n \to 1$. Since e^x is continuous, it follows that $e^{\ln(1+\frac{1}{n})^n} = \{(1+\frac{1}{n})^n\} \to e^1 = e$. □

Solution to Exercise 6.9. *Let $f, g : [a,b] \to \mathbb{R}$ be integrable functions. Then fg is bounded.*

Proof. Since f, g are integrable, they are both bounded functions. Thus, we may choose $M, N > 0$ so that $|f(x)| \leq M$ and $|g(x)| \leq N$. Then $|f(x)g(x)| = |f(x)||g(x)| \leq MN$.

□

Solution to Exercise 6.10. *Define $\frac{e^x + e^{-x}}{2} = \cosh(x)$ and $\frac{e^x - e^{-x}}{2} = \sinh(x)$. Then $(\sinh(x))' = \cosh(x)$ and $(\cosh(x))' = \sinh(x)$.*

Proof. We just use the linearity of the integral and the chain rule to get $(\frac{e^x + e^{-x}}{2})' = \frac{e^x - e^{-x}}{2}$ and $(\frac{e^x - e^{-x}}{2})' = \frac{e^x + e^{-x}}{2}$. □

Solution to Exercise 6.11. *Find, with proof, an example of a function $f(x)$ which is integrable on $[a,b]$ so that $F(x) = \int_a^x f(t)dt$ is not differentiable.*

Proof. Let $f(x) = 0$ if $0 \leq x \leq 1$ and let $f(x) = 1$ if $1 \leq x \leq 2$ and let $F(x) = \displaystyle\int_0^x f(t)dt$. We know f is integrable because it has only one discontinuity, so f is integrable by the Lebesgue Characterization of Riemann Integrability. However, $\displaystyle\lim_{x\to 1^-} \frac{F(1) - F(x)}{1-x} = \lim_{x\to 1^-} \frac{\int_x^1 0dt}{1-x} = 0$, whereas $\displaystyle\lim_{x\to 1^+} \frac{F(x) - F(1)}{x-1} = \lim_{x\to 1^+} \frac{\int_1^x 1dt}{x-1} = \lim_{x\to 1^+} \frac{x-1}{x-1} = 1$. Thus, F is not differentiable at $x = 1$. □

Solution to Exercise 6.12. *Prove that if $c > 0$ and $a, b \in \mathbb{R}$ then:*

(a) $c^a c^b = c^{a+b}$.

(b) $\dfrac{c^a}{c^b} = c^{a-b}$

(c) $(c^a)^b = c^{ab}$.

Proof. (a) We have shown in Theorem 6.25 that $\ln(c^a c^b) = c\ln(a) + b\ln(c) = (a+b)\ln(c) = \ln(c^{a+b})$. Since $\ln(x)$ is one to one, it follows that $c^a c^b = c^{a+b}$.

(b) We know from Theorem 6.25 that $\ln(\dfrac{c^a}{c^b}) = a\ln(c) - b\ln(c) = (b-c)\ln(c) = \ln(c^{b-a})$. Since $\ln(x)$ is one to one, we see that $\dfrac{c^a}{c^b} = c^{a-b}$.

(c) We know from Theorem 6.25 that $\ln((c^a)^b) = b\ln(c^a) = ab\ln(c) = \ln(c^{ab})$. Since $\ln(x)$ is one to one this implies that $(c^a)^b = c^{ab}$. □

Solution to Exercise 6.13. *Prove that if $\{x_n\}$ is a sequence of positive numbers converging to a real number r and $c > 0$ then $\{c^{x_n}\} \to c^r$.*

Proof. Since c^x is continuous since it is differentiable by Theorem 6.26, by the Sequential Characterization of Continuity, $\{c^{x_n}\} \to c^r$. □

Solution to Exercise 6.14. *Let $f : [a, b] \to \mathbb{R}$ be integrable. Then $|f|$ is integrable, and $|\displaystyle\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx$.*

Proof. We know $|f|$ is integrable by Theorem 6.10 since f is integrable and $|x|$ is continuous. Since $-|f(x)| \leq f(x) \leq |f(x)|$, we know that $-\displaystyle\int_a^b |f(x)|dx \leq \int_a^b f(x)dx \leq \int_a^b |f(x)|dx$ by Theorem 6.4, which means that $|\displaystyle\int_a^b f(x)dx| \leq \int_a^b |f(x)|dx$ by Theorem 1.16.

□

Chapter 7

Supplementary Materials for One Variable

7.1 The Natural Numbers

Definition 42

A set $S \subseteq \mathbb{R}$ is *inductive* if $1 \in S$ and for every $k \in S$ it is true that $k+1 \in S$. The set of *natural numbers* $\mathbb{N}$ is the intersection of all inductive sets. A set S is *well-ordered* if every non-empty subset of S has a least element.

One example of an inductive set is $\mathbb{R}$, so we know that inductive sets exist.

Theorem 7.1. *(a) $\mathbb{N}$ is an inductive set*

(b) 1 is the least element of $\mathbb{N}$. Furthermore, $S = \{1\} \cup [2, \infty)$ is inductive.

(c) If $n > 1$ and $n \in \mathbb{N}$ then $n - 1 \in \mathbb{N}$

Proof. (a) Since 1 is an element of all inductive sets (and there are inductive sets that exist), $1 \in \mathbb{N}$. If $k \in \mathbb{N}$ then for every inductive set S it follows that $k \in S$ and thus $k+1 \in S$. Hence $k+1 \in \mathbb{N}$ and so $\mathbb{N}$ is inductive.

(b) We know that $1 \in S = \{1\} \cup [2, \infty)$, and if $x < 2$ and $x \in S$ then $x = 1$ so $x + 1 = 2 \in S$. If $x \geq 2$ then $x + 1 > 2$ so $x + 1 \in [2, \infty)$ so $x + 1 \in S$. Thus, S is inductive, so $\mathbb{N} \subseteq S$ and since 1 is the least element of S it follows that $\mathbb{N}$ contains no points that precede 1.

(c) Suppose there is a natural number $l > 1$ so that $l - 1 \notin \mathbb{N}$. Let $W = \mathbb{N} \setminus \{l\}$. Then since $l > 1$ we know that $1 \in W$. Likewise, for any $x \in W$ we know that $x + 1 \in W$ since $x + 1 \in \mathbb{N}$ and $x + 1 \neq l$. Hence W is inductive and does not contain l, which is impossible since $\mathbb{N}$ only contains numbers which are elements of every inductive set. We conclude that if $n > 1$ and $n \in \mathbb{N}$ then $n - 1 \in \mathbb{N}$.

□

Theorem 7.2. *Principle of Mathematical Induction. For each $n \in \mathbb{N}$, let $P(n)$ be a statement so that (a) $P(1)$ is true, and (b) whenever $P(k)$ is true for a natural number k, it follows that $P(k+1)$ is also true. Then it follows that $P(n)$ is true for all $n \in \mathbb{N}$.*

Proof. Let $S = \{n \in \mathbb{N} | P(n) \text{ is true }\}$. Then by (a) we know that $1 \in S$ and by (*b*) we know that if $k \in S$ then $k + 1 \in S$. Therefore S is inductive, so $\mathbb{N} \subseteq S$, which means that $P(n)$ is true for every $n \in \mathbb{N}$.

□

Theorem 7.3. *(a) If $n \in \mathbb{N}$ then there are no elements of $\mathbb{N}$ between n and $n+1$ or between $n - 1$ and n.*

(b) The natural numbers are well-ordered.

Proof. (a) Let $P(n)$ be the statement that there are no natural numbers between j and $j + 1$ for all natural numbers $j \leq n$. Since $\{1\} \cup [2, \infty)$ is inductive by Theorem 7.1 we know that this set contains $\mathbb{N}$ which means that there are no natural numbers between 1 and 2, so $P(1)$ is true. Assume that there are no natural numbers between i and $i + 1$ for all $i \leq k - 1$ for some $k \in \mathbb{N}$. Then let $S = \mathbb{N} \cap ([1, k] \cup [k + 1, \infty))$. If $x \in \mathbb{N} \cap [1, k - 1]$ then $x + 1 \leq k$ and $x + 1 \in \mathbb{N}$ since $\mathbb{N}$ is inductive, which means $x + 1 \in S$. Since there are no natural numbers between $k - 1$ and k, for any $x \in \mathbb{N} \cap [1, k]$, we know that either $x \in [1, k-1]$ in which case $x+1 \in S$, or $x = k$ in which case $k-1+1 = k \in S$, so $x+1 \in S$. Also, if $x \geq k$ then $x + 1 \in \mathbb{N}$ (since $\mathbb{N}$ is inductive) and $x + 1 \in [k + 1, \infty)$, which means $x + 1 \in S$. Thus, S is inductive so $\mathbb{N} \subseteq S$, which means that there are no natural numbers between k and $k + 1$. By induction, it follows that for each $n \in \mathbb{N}$, there are no elements of $\mathbb{N}$ between n and $n + 1$.

We know there are no elements of $\mathbb{N}$ in $(0, 1)$ (since 1 is the least natural number), and if $m > 1$ and $m \in \mathbb{N}$ then $m - 1 \in \mathbb{N}$ by Theorem 7.1, so there are no elements of $\mathbb{N}$ in $(m - 1, m)$. Hence, for each natural number n it is true that there are no natural numbers between $n - 1$ and n.

(b) For each natural number n, let $P(n)$ be the statement: If S is a subset of the natural numbers that intersects the set of all natural numbers less than or equal to n then S has a least element. We note that $P(1)$ is true since if S contains 1 then 1 is the least element of S since 1 is the least natural number. Next, assume $P(k)$ is true for $k \in \mathbb{N}$. Let S be a subset of $\mathbb{N}$ which intersects the set of natural numbers less than or equal to $k + 1$. If S contains a natural number less than or equal to k then S has a least element by the induction hypothesis. If S does not contain any natural numbers less than or equal to k then by (a) we know that there are no natural numbers between k and $k + 1$ which means that the least element of S is $k + 1$. The result follows for all $n \in \mathbb{N}$ by induction. Since any non-empty subset S of the natural numbers must be a subset of the natural numbers that intersects the set of all natural numbers less than or equal to one of the elements of S, it follows that every non-empty subset of $\mathbb{N}$ has a least element, so S is well ordered.

□

Definition 43

We will let $\{1, 2, 3, ..., n\}$ be notation to denote $\mathbb{N} \cap [1, n]$ for any natural number n. The *n-tuple* $(x_1, x_2, x_3, ...x_n)$ is an ordered set of n real numbers which is a function $f : \{1, 2, 3, ..., n\} \to \mathbb{R}$, where $f(i) = x_i$ for each $i \in \{1, 2, 3, ..., n\}$.

Theorem 7.4. *Recursive Definition.*

For each natural number n we let $P_n(z_1, z_2, ..., z_n)$ be a statement about n real numbers listed in order and satisfying the following criteria:

(1) $P_1(x)$ is true for exactly one point $x = x_1 \in \mathbb{R}$.

(2) If n is a natural number greater than one then for any $x_1, x_2, x_3, ..., x_{n-1}$ so that $P_{n-1}(x_1, ..., x_{n-1})$ is true there is exactly one x_n so that $P_n(x_1, ..., x_{n-1}, x_n)$ is true.

Then for each $n \in \mathbb{N}$ there is a unique function $f_n : \{1, 2, 3, ..., n\} \to \mathbb{R}$ so that, if we define $x_i = f_n(i)$ for each $i \in \{1, 2, 3, ..., n\}$, then $P_n(x_1, ..., x_n)$ is true,

and $f_n(i) = f_m(i)$ for all natural numbers $i \leq m \leq n$. Furthermore, there is a unique function $f : \mathbb{N} \to \mathbb{R}$ so that $P_n(x_1, x_2, ..., x_n)$ is true for each $n \in \mathbb{N}$.

Proof. We proceed by induction to show that for each natural number n there is a unique $f_n : \{1, 2, ..., n\} \to \mathbb{R}$ so that $P_n(x_1, ..., x_n)$ is true and $f_n(i) = f_m(i)$ for all natural numbers $i \leq m \leq n$. This is true if $n = 1$ by (1). Assume this is true if $n = k$. Then there is a unique $f_k : \{1, 2, 3, ..., k\} \to \mathbb{R}$ so that $P_k(x_1, x_2, x_3, ..., x_k)$ is true and $f_k(i) = f_j(i)$ for all natural numbers $i \leq j \leq k$. By (2) there is a unique x_{k+1} so that $P_{k+1}(x_1, ..., x_k, x_{k+1})$ is true by hypothesis, so there is only one choice for f_{k+1} meeting the requirements that $f_{k+1}(i)$ must equal $f_k(i)$ for all natural numbers $i \leq k$ and $P_{k+1}(x_1, ..., x_k, x_{k+1})$ is true, namely $f_{k+1}(k+1) = x_{k+1}$ and $f_{k+1}(i) = f_k(i)$ for all natural numbers $i \leq k$. Thus, $f_{k+1}(i) = f_j(i)$ for all natural numbers $i \leq j \leq k+1$ since $f_j(i) = f_k(i)$ for $i \leq k$.

If we define $f : \mathbb{N} \to \mathbb{R}$ by $f(n) = f_n(n)$ for each $n \in \mathbb{N}$ then by the argument above we know that $P_n(x_1, ..., x_n)$ is true for all $n \in \mathbb{N}$. Since $(x_1, x_2, ..., x_n)$ is the unique n-tuple so that $P_i(x_1, ..., x_i)$ is true for all $1 \leq i \leq n$, it follows that the choice of f is unique.

□

The process of generating a sequence by recursive definition is the process described here. Generally, we start with one or more numbers at the beginning of such a sequence and then assign a rule to define the next number in the sequence based on its predecessors. Initially, when talking about $\mathbb{N}$, we would have liked to simply say that $\mathbb{N} = \{1, 2, 3, 4, ...\}$ which was meaningless because it depended on an undefined notion of dots at the end. Now, we can instead say that we would like to define statement $P(1)$ to be $x_1 = 1$, which is true if any only if $x_1 = 1$. Then we can say that if $n \in \mathbb{N}$ and $n > 1$ then we define

$P(n)$ to be the statement that $x_n = x_{n-1} + 1$. If we have already established the n-tuple $(x_1, x_2, ..., x_{n-1})$ so that $P_i(x_1, ..., x_i)$ is true for all $i \leq n - 1$ then there is only one x_n satisfying P_n. In other words, once we know x_{n-1}, the assignment $x_n = x_{n-1} + 1$ is unique. Hence, we have established that the function $f : \mathbb{N} \to \mathbb{R}$ induced by these statements is unique. We can inductively argue that $f(n) = n$ for each natural number n under this mapping. First, $f(1) = 1$, and if $f(k) = k$ then by definition $f(k + 1) = f(k) + 1 = k + 1$. This means that the set defined by the rule "start at one, and then for each element of the set the element obtained by adding one to that element is the next element of the set and continue indefinitely" is the set of natural numbers. This is what is actually meant by $\mathbb{N} = \{1, 2, 3, ...\}$ since the "..." is intended to mean "continue with the rule that the next element of the set is the preceding element plus one." Thus, now that we have proven the preceding theorem, it is reasonable and correct for us to write $\mathbb{N} = \{1, 2, 3, ...\}$ if the rule choosing the unique next element in the list is understood and is clearly unique based on its predecessors and a unique starting point is included, and the statement now has meaning. Recall, though, that the domain of the function f used to define the natural numbers in this way has a domain which is the natural numbers. In other words, we had to establish the properties of the natural numbers before we could use this description of the natural numbers.

It should also be understood that the statement may vary depending on a choice for recursive definition, in which case the sequence is unique based on those choices, but the choices are usually unknown so which sequence is determined by those choices is not known (there would be different possibilities for different choices). For example, you could say pick $x_1 \in \mathbb{R}$ to be the first member of a sequence, let x_2 be a number more than x_1 and then choose a number $x_3 > x_2$ and so on. After the choice of x_1 is made (but not before) the statement that "x_1 is the point that was chosen" is satisfied by only one real number. The sequence chosen is an increasing sequence and exists. The specific property each choice satisfies based on the preceding statements is that the next choice is chosen from among those points that are greater than the preceding choice. There were many choices possible, and the choice was not unique (but the point chosen became unique once the choice was made, and was the unique point satisfying the statement that the point was the one which was selected). Nevertheless, that is a legitimate way of creating a sequence (even though its members are not specified if the choices were never stated) to make choices among options at each stage for the next element in the sequence.

7.2 Fundamental Theorem of Arithmetic

Theorem 7.5. *Euclid's Division Lemma. Let $a \in \mathbb{Z}$ and let $b \in \mathbb{N}$. Then there are unique integers q and r so that $a = bq + r$ and $0 \leq r < b$.*

Proof. Let $S = \{a - nb \in \mathbb{N} \cup \{0\} | n \in \mathbb{Z}\}$. Then S is non-empty since by Theorem 2.6, we know that there is a natural number $m > \dfrac{-a}{b}$, so $a - (-m)b > 0$. Hence, since $\mathbb{N}$ is well-ordered, $S \cap \mathbb{N}$ has a smallest element, which means that S has a smallest element (which is zero if $0 \in S$ and the smallest element of $S \cap \mathbb{N}$ otherwise). We write this element as $a - qb = r$ for some $q \in \mathbb{Z}$. Since $a - qb$ is the least element of S it must follow that $a - qb < b$ or else $a - (q + 1)b \geq b - b = 0$, so $a - (q + 1)b \in S$, contradicting $a - qb$ being the least element of S. Since $0 \leq a - qb < b$ we know that $a = bq + r$ and $0 \leq r < b$.

To see that the choice of q and r is unique, first note that for any integers Q, R so that $a = bQ + R$ it must follow that $R = a - bQ$, so R is uniquely determined by Q. If $Q > q$ then $R = a - bQ \leq a - (q+1)b = r - b < 0$, and if $Q < q$ then $R = a - bQ \geq a - (q-1)b = r + b \geq b$. Hence, there is only one possible choice of q which makes $a = bq + r$ and $0 \leq r < b$ for a uniquely associated integer r.

□

Definition 44

Let $a \in \mathbb{Z}$ and let $b \in \mathbb{N}$. Let q and r be the unique integers so that $a = bq + r$ and $0 \leq r < b$. Then we say that q is the *quotient* when a is divided by b and that r is the *remainder*.

Theorem 7.6. *Bezout's Identity. Let x and y be integers. Then the smallest positive integer of the form $ix + jy$, where $i, j \in \mathbb{Z}$, is the greatest common divisor of x and y. Furthermore, the integers of the form $ix + jy$, where $i, j \in \mathbb{Z}$, are the integer multiples of d.*

Proof. Let $S = \{mx + ny \in \mathbb{N} | m, n \in \mathbb{Z}\}$. Then S is non-empty (either $x + y \in \mathbb{N}$ or $-x - y \in \mathbb{N}$) so since $\mathbb{N}$ is well-ordered we know that S has a least element $d = sx + ty$ for some integers s, t. Using Euclid's Division Lemma, we know that there are unique inters q, r so that $x = qd + r$ and $0 \leq r < d$. This means that $r = x - qd = x - q(sx + ty) = (1 - qs)x - qty \in S \cup \{0\}$. However, since $r < d$ and d is the smallest element of S we know that $r = 0$. From this we conclude that $x = qd$. We can similarly show that for some integer w it is true that $y = wd$, which means that d is a divisor of x and y. If c is any other positive divisor of x and y then there are integers c_x, c_y to that $x = c_x c$ and $y = c_y c$. That means that $d = sc_x c + tc_y c = c(sc_x + tc_y)$, which means that $sc_x + tc_y \geq 1$ (since multiplying a negative number or zero by a positive integer cannot result in a positive integer). Thus $c(sc_x + tc_y) \geq c(1)$, so $d \geq c$, meaning that d is the greatest common divisor.

Finally, for any integer of the form $ix + jy$, we know that $ix + jy = iqd + jwd = d(iq + jw)$ is a multiple of d. Likewise, any multiple of d by an integer k is equal to $kd = k(sx + ty) = (ks)x + (kt)y$, which is an integer of the form $ix + jy$. □

Theorem 7.7. *Euclid's Lemma. Let a, b be integers and let $p \in \mathbb{N}$ so that the greatest common divisor of p and a is one and $p|ab$. Then $p|b$.*

Proof. By Bezout's Identity, we know that we can find integers s, t so that $sp + ta = 1$ which means that $bsp + bta = b$. Since $p|ab$ we can find an integer m so that $pm = ab$. Hence, $psb + pmt = b$ so $p(sb + tm) = b$, which means that p divides b.

□

Theorem 7.8. *Corollary to Euclid's Lemma. Let p be a prime number which divides q^k, where q is a prime number and k is a positive integer. Then $p = q$. Furthermore, the greatest common divisor of one prime and another prime taken to a positive power is always one.*

Proof. Suppose $p \neq q$. Then p does not divide q since q is prime, so the greatest common divisor of p and q is one. Hence, if $p|q^2$ then $p|q$ by Euclid's Lemma, which is impossible, which means that the greatest common divisor of p and q^2 is one. Proceeding inductively, if we have shown that the greatest common divisor of q^r and p is one for some positive integer r then if $p|q^{r+1}$ then $p|q^r q$, so it follows from Euclid's Lemma that $p|q$ which it cannot. Thus, p does not divide q^{r+1} and therefore the greatest common divisor of p and q^{r+1} is one. Thus p cannot divide any positive power of q and has greatest common divisor of one with every positive power of q. □

Definition 45

Let $\{p_1, p_2, ..., p_k\}$ be a set of prime numbers and $\{m_1, m_2, ..., m_k\} \subset \mathbb{N}$. If $n = p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_k^{m_k}$ then we refer to $p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_k^{m_k}$ as a *prime factorization* or *prime decomposition* for n (or of n).

In this definition we said "a" prime factorization, but we are about to prove there is only one prime decomposition for any positive integer.

Theorem 7.9. *The Fundamental Theorem of Arithmetic. Let $n > 1$ where $n \in \mathbb{N}$. Then there is a unique set of prime numbers $\{p_1, p_2, ..., p_k\}$ with corresponding positive integer powers $\{m_1, m_2, ..., m_k\}$ so that $n = p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_k^{m_k}$.*

Proof. We first show there is such a set of powers of prime numbers. We proceed by strong induction on n. First, if $n = 2$ then n can be represented as $n = 2^1$. Assume that we have shown there is a representation of n as a product of prime numbers for all $n \leq k$. If $k+1$ is prime then $k+1 = (k+1)^1$. If $k+1$ is composite then $k+1 = st$, where s, t are positive integers greater than one and thus also less than $k+1$ (since a number greater than one times a number greater than or equal to $k+1$ is greater than $k+1$). By the induction hypothesis, since $1 < s \leq k$ and $1 < t \leq k$ there is a representation of s, t as a product of prime numbers raised to positive integer powers and we can write $s = p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_u^{m_u}$, $t = q_1^{j_1} q_2^{j_2} q_3^{j_3} ... q_v^{j_v}$, where each p_i and q_i is prime and each m_i and j_i is a positive integer. Thus, $k+1 = st = p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_u^{m_u} q_1^{j_1} q_2^{j_2} q_3^{j_3} ... q_v^{j_v}$ is a product of primes raised to positive powers.

To show uniqueness, suppose $S = \{n \in \mathbb{N} | n \text{ has two different prime factorizations }\}$ is non-empty. Then since $\mathbb{N}$ is well ordered, there is a least integer $n \in S$. Then there are factorizations $n = p_1^{m_1} p_2^{m_2} p_3^{m_3} ... p_u^{m_u}$ and $n = q_1^{j_1} q_2^{j_2} q_3^{j_3} ... q_v^{j_v}$ for n (where each p_i and q_i is prime and each m_i and j_i is a positive integer). Since $p_1 | n$, by Euclid's Lemma we know that p_1 divides some $q_k^{j_k}$ and therefore $p_1 = q_k$ for some positive integer k by the Corollary to Euclid's Lemma (since p_1, q_k are prime). Thus, $\frac{n}{p_1}$ is an integer less than n which can be represented with two different prime factorizations, namely $\frac{n}{p_1} = p_1^{m_1 - 1} p_2^{m_2} p_3^{m_3} ... p_u^{m_u} = q_1^{j_1} q_2^{j_2} q_3^{j_3} ... q_k^{j_k - 1} ... q_v^{j_v}$. This contradicts the assumption that n was the smallest element of S. We conclude that there is exactly one prime factorization of any natural number greater than one. □

Theorem 7.10. *Let $\frac{m}{n} \in \mathbb{Q}$, where $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. Then $\frac{m}{n} = \frac{p}{q}$ for some $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, where the greatest common divisor of p and q is one.*

Proof. Since $\mathbb{N}$ is well ordered, there is a smallest natural number q so that $\frac{p}{q} = \frac{m}{n}$ for some integer p. Given q, p is uniquely determined, since $p = \frac{qm}{n}$. If k is an integer greater than one which is a common divisor of q and p then we can write $\frac{p}{q}$ as $\frac{sk}{tk}$ for integers s, t, which means that $t < q$ and it is possible to write $\frac{m}{n} = \frac{s}{t}$, contradicting our choice of q. Hence, the greatest common divisor of p and q is one. □

Definition 46

Let $\frac{p}{q} \in \mathbb{Q}$ for some $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. We say $\frac{p}{q}$ is written in *reduced terms* if p and q are chosen so that the greatest common divisor of p and q is one.

Theorem 7.11. *The square root of a prime number p is irrational.*

Proof. We know that the square root of every prime number exists by Exercise 4.12. Suppose $\sqrt{p} = \frac{m}{n}$ for some $m \in \mathbb{Z}$ and $n \in \mathbb{N}$ written in reduced terms. Then $n^2 p = m^2$ which means that p is a divisor of m^2 and is therefore in the prime decomposition of m^2, which is only possible if p is in the prime decomposition of m by the Fundamental Theorem of Arithmetic, which means that m^2 is divisible by p^2. But this means $\frac{m^2}{p} = n^2$ is divisible by p, so n is divisible by p, contradicting the assumption that $\frac{m}{n}$ was written in reduced terms. We conclude that $\sqrt{p}$ is irrational.

□

7.3 Decimals

Definition 47

If the sequence $\{d_0, d_0+\frac{d_1}{10}, d_0+\frac{d_1}{10}+\frac{d_2}{100}+...\} \to r$, where $d_0 \in \mathbb{N}$ and $0 \leq d_i \leq 9$ for each $i \in \mathbb{N}$, then we say that $d_0.d_1d_2d_3...$ is a *decimal expansion* for r and that $r = d_0.d_1d_2d_3....$ We will refer to a sequence $\{d_0, d_0+\frac{d_1}{10}, d_0+\frac{d_1}{10}+\frac{d_2}{100}, ...\}$, where $d_0 \in \mathbb{N}\cup\{0\}$ and $0 \leq d_i \leq 9$ for each $i \in \mathbb{N}$ as a *positive decimal sequence* (even though it might consist entirely of zero entries and not actually converge to a positive number). We refer to $d_0.d_1d_2...d_n = d_0, d_0+\frac{d_1}{10}, d_0+\frac{d_1}{10}+\frac{d_2}{100}+...+\frac{d_n}{10^n}$ as a *terminating decimal expansion* and note that this is the same as $d_0.d_1d_2...d_n000...$ (since this is a constant sequence after the nth term and thus converges). We also refer to $\{-d_0, -d_0-\frac{d_1}{10}, -d_0-\frac{d_1}{10}-\frac{d_2}{100}, ...\}$ as a *negative decimal sequence.* If $\{-d_0, -d_0-\frac{d_1}{10}, -d_0-\frac{d_1}{10}-\frac{d_2}{100}, ...\} \to r$ then we say that $r = -d_0.d_1d_2d_3...$, and $-d_0.d_1d_2d_3...$ is a decimal expansion for r.

We will refer to the *primary* decimal representation of a non-negative real number r as $d_0.d_1d_2d_3...$ if d_0 is the greatest non-negative integer which is less than or equal to r, d_1 is the largest non-negative integer so that $\frac{d_1}{10}$ is less than or equal to $r-d_0$ (so $0 \leq d_1 \leq 9$ since $r-d_0 < 1$), and d_2 is the largest non-negative integer so that $\frac{d_2}{100}$ less than or equal to $r-d_0-\frac{d_1}{10}$ (so $0 \leq d_2 \leq 9$), and so on, in which case $\{d_0, d_0.d_1, d_0.d_1d_2, ...\}$ is the *primary decimal sequence* for r.

If r is a negative real number and the primary decimal representation for $-r$ is $d_0.d_1d_2d_3...$, then we write the primary decimal representation of $-r$ as $-d_0.d_1d_2d_3...$, and the negative decimal sequence $\{-d_0, -d_0-\frac{d_1}{10}, -d_0-\frac{d_1}{10}-\frac{d_2}{100}, ...\}$ is the primary decimal sequence for $-r$.

Theorem 7.12. *If $d_0.d_1d_2d_3...$ is the primary decimal representation for a non-negative real number r then $r = d_0.d_1d_2d_3....$ Likewise, if $-d_0.d_1d_2d_3...$ is the primary decimal representation for a negative number $-r$ then $-r = -d_0.d_1d_2d_3...$*

Proof. We first observe inductively that, for all positive integers n, it is true that $r - d_0.d_1d_2...d_n < \frac{1}{10^n}$. This is true for $n = 1$ by definition. Assuming this is true for $n = k$, so $r - d_0.d_1d_2...d_k < \frac{1}{10^k}$. Recall that d_{k+1} is chosen to be the largest non-negative integer so that $r - d_0.d_1d_2...d_k - \frac{d_{k+1}}{10^{k+1}} \geq 0$. It must follow that $r - d_0.d_1d_2...d_k - \frac{d_{k+1}+1}{10^{k+1}} < 0$, which means that $r - d_0.d_1d_2...d_k - \frac{d_{k+1}}{10^{k+1}} < \frac{1}{10^{k+1}}$. We conclude that $r - d_0.d_1d_2...d_n < \frac{1}{10^n}$ for each $n \in \mathbb{N}$. By Exercise 3.13, we know that $\{\frac{1}{10^n}\} \to 0$, so since $r - \frac{1}{10^n} \leq r - d_0.d_1d_2...d_n < r + \frac{1}{10^n}$, by the Squeeze Theorem we see that the primary decimal sequence

$\{d_0, d_0 + \frac{d_1}{10}, d_0 + \frac{d_1}{10} + \frac{d_2}{100} + ...\} \to r$, so $r = d_0.d_1d_2....$

Having shown that $\{d_0, d_0 + \frac{d_1}{10}, d_0 + \frac{d_1}{10} + \frac{d_2}{100} + ...\} \to r$, it follows that $\{-d_0, -d_0 - \frac{d_1}{10}, -d_0 - \frac{d_1}{10} - \frac{d_2}{100} + ...\} \to -r$ as well.

□

It is not always true that a decimal expansion for a number is unique. For instance, 0.99999... = 1.00000. To verify this, just note that 0.9999...9 with n entries of p is the same as $1 - \frac{1}{10^{n+1}}$ and $\{1 - \frac{1}{10^{n+1}}\} \to 1$ since we know $\{\frac{1}{10^{n+1}}\} \to 0$ by Exercise 3.13. Likewise, $\{\frac{1}{10^k}(1 - \frac{1}{10^{n+1}})\} \to \frac{1}{10^k}$, which means that 0.000...0999... = 0.000...1, where the first k entries on the left hand side are zero after the decimal point and the kth entry after the decimal point on the right is 1 (and the other entries are zero).

Note that all the theorems proven about $\mathbb{N}$ in Chapter 2 follow without the completeness axiom, apart from the theorem that $\mathbb{N}$ is not bounded above (which need not be true if the the least upper bound property is false). This is important because some decimal properties are established using induction, and hold whether or not the completeness axiom is assumed.

Theorem 7.13. *Let $r = d_0.d_1d_2...$ be the limit of a positive decimal sequence $\{d_0, d_0 + \frac{d_1}{10}, d_0 + \frac{d_1}{10} + \frac{d_2}{100}, ...\}$ in an ordered field S. Then $d_0.d_1d_2d_3...d_n \leq r \leq d_0.d_1d_2...d_n + \frac{1}{10^n}$ for each natural number n. Furthermore, if $-r = -d_0.d_1d_2d_3...$ is the limit of a negative decimal sequence $\{-d_0, -d_0 - \frac{d_1}{10}, -d_0 - \frac{d_1}{10} - \frac{d_2}{100}, ...\}$ then $-d_0.d_1d_2d_3...d_n - \frac{1}{10^n} \leq r \leq -d_0.d_1d_2d_3...d_n$.*

Proof. By the Comparison Theorem (which did not require the completeness axiom to prove), we know that since all terms of the positive decimal sequence $\{d_0, d_0 + \frac{d_1}{10}, d_0 + \frac{d_1}{10} + \frac{d_2}{100}, ...\}$ after the nth term are at least as large as $d_0.d_1d_2d_3...d_n$, it follows that $r \geq d_0.d_1d_2d_3...d_n$. Similarly, for any natural number k, we notice that $d_0.d_1d_2d_3...d_nd_{n+1}...d_{n+k} = d_0.d_1d_2d_3...d_n + \frac{1}{10^n}(0.d_{n+1}d_{n+2}...d_{n+k}) \leq d_0.d_1d_2d_3...d_n + \frac{1}{10^n}(0.999...) = d_0.d_1d_2d_3...d_n + \frac{1}{10^n}$. Thus, by the Comparison Theorem again, we get that $d_0.d_1d_2d_3...d_n \leq r \leq d_0.d_1d_2...d_n + \frac{1}{10^n}$. From this it also follows that $-d_0.d_1d_2d_3...d_n - \frac{1}{10^n} \leq -r \leq -d_0.d_1d_2d_3...d_n$.

□

The Monotone Convergence Theorem is used in part of the proof below. Recall that the completeness axiom was necessary to prove that theorem.

Theorem 7.14. *Let F be an ordered field. Then every set in F which is bounded above has a least upper bound if and only if the set $\mathbb{N}$ of natural numbers is not bounded above and every decimal sequence converges.*

Proof. First, assume that all decimal sequences converge. Let S be a non-empty subset of F which is bounded above. For now, assume that $S \cap [0, \infty) \neq \emptyset$. Since $\mathbb{N}$ is not bounded above, there is a first natural number $d_0 + 1$ which exceeds all elements of S since $\mathbb{N}$ is well ordered, which means that d_0 is the largest integer which does not exceed all elements of S. Since $d_0 + 1$ does exceed all elements of S, there is a largest integer d_1 so that $0 \leq d_1 \leq 9$ so that $d_0 + \dfrac{d_1}{10}$ does not exceed all elements of S. If we have chosen d_i for all $i \leq k \in \mathbb{N}$ so that $d_0.d_1d_2...d_k$ does not exceed all elements of S and for each i it is true that $d_0.d_1d_2..d_i + \dfrac{1}{10^i}$ does exceed all elements of S, then we know $d_0.d_1d_2...d_k + \dfrac{1}{10^k}$ does exceed all elements of S, so we choose $0 \leq d_{k+1} \leq 9$ to be the largest integer so that $d_0.d_1d_2...d_kd_{k+1}$ does not exceed all elements of S. This lets us inductively choose d_n for all $n \in \mathbb{N}$. We claim that $r = d_0.d_2d_2...$ is the least upper bound for S.

To see that r is an upper bound for S, suppose there is some $s \in S$ so that $s > r$. Then $s - r > \dfrac{1}{10^k}$ for some positive integer k by Exercise 3.13. But we know that $d_0.d_1d_2...d_k + \dfrac{1}{10^k}$ exceeds all elements of S so $d_0.d_1d_2...d_k \leq r < s < d_0.d_1d_2...d_k + \dfrac{1}{10^k}$, so this is impossible. To see that r is the least upper bound for S, let $u < r$. Then $r - u > 0$ and we can, again, choose k so that $\dfrac{1}{10^k} < r - u$. We know that $d_0.d_1d_2...d_k$ does not exceed all elements of S and $r - d_0.d_1d_2...d_k \leq \dfrac{1}{10^k}$ by Theorem 7.13, so there is some $t \in S$ so that $t \geq d_0.d_1d_2...d_k$, and $r - t \leq \dfrac{1}{10^k} < r - u$, so $t > u$, which means that u is not an upper bound for S, and therefore r is the least upper bound of S.

If S contains only negative numbers, then S contains a negative number k and so $(S - k) \cap [0, \infty) \neq \emptyset$ and thus $S - k$ has a least upper bound r. For any $x \in S$, it follows that $x - k \leq r$, so $x \leq r + k$, so $r + k$ is an upper bound for S. If $u < r + k$ then $u - k < r$, so $u - k$ is not an upper bound for $S - k$, and there is some $x - k > u - k$ for some $x \in S$. This means that $x > u$, so u is not an upper bound for S. Hence, $\sup(S) = r + k$.

Next, assume the completeness axiom. Then let $\{d_0, d_0.d_1, d_0.d_1d_2, ...\}$ be a positive decimal sequence. The sequence is non-decreasing since each term is the preceding term plus a non-negative number. By Theorem 7.13, we know that each $d_0.d_1d_2...d_n \leq d_0 + 1$, which means that this decimal sequence is a non-decreasing sequence which is bounded above and therefore converges to a real number r by the Monotone Convergence Theorem.

Similarly, for any negative decimal sequence $\{-d_0, -d_0 - \dfrac{d_0}{10}, ...\}$ we know the corresponding positive decimal sequence $\{d_0, d_0.d_1, d_0, d_1d_2, ...\}$ converges to some number r, so the negative decimal sequence converges to $-r$. Hence, assuming the completeness axiom we know that all decimal sequences converge to real numbers.

The fact that the natural numbers is not bounded above also follows from the completeness axiom, and was proven in Theorem 2.6.

□

We conclude with a theorem that will make the decimal-based argument that the real numbers are uncountable that we will discuss in the cardinality portion of the chapter rigorous.

Theorem 7.15. *Let $r = r_0.r_1r_2r_3...$, $s = s_0.s_1s_2s_3...$ be any non-negative real numbers so that for some non-negative integer i it is the case that $r_i \neq s_i$, and m is the first non-negative integer so that $r_m \neq s_m$. If $r_m < s_m$ then $r < s$ unless $s_m = r_m + 1$, in which case $r = s$ if and only if $r_i = 9$ for all $i > m$ and $s_i = 0$ for all $i > m$. In particular, if $1 < |r_i - s_i| < 9$ for some positive integer i then $r \neq s$.*

Proof. Without loss of generality we may assume that $r_m < s_m \leq 9$ as described above. We know that $r_0.r_1r_2...r_m < r_0.r_1r_2...r_m + \dfrac{1}{10^m} \leq s_0.s_1s_2s_3...s_m \leq s$ by Theorem 7.13. Additionally, by the Comparison Theorem, we note that if any $s_j > 0$ for $j > m$ then $s_0.s_1s_2s_3...s_m < s_0.s_1s_2s_3...s_j \leq s$, so $r < s$. Likewise, if $r_j < 9$ for any $j > m$ then $r \leq r_0.r_1r_2...r_j + \dfrac{1}{10^j} = r_0.r_1r_2...r_{j-1}(r_j + 1) \leq r_0.r_1r_2..r_m999...9(r_j + 1) \leq r_0.r_1r_2..r_m + \dfrac{1}{10^m}(1 - \dfrac{1}{10^{j-m}}) < s_0.s_1s_2s_3...s_m \leq s$. Hence, it is only possible for r and s to be equal if $s_j = 0$ for $j > m$ and $r_j = 9$ for all $j > m$, in which case $r = r_0.r_1r_2...r_m + \dfrac{1}{10^m}(0.999...) = r_0.r_1r_2...r_m + \dfrac{1}{10^m} = r_0.r_1r_2...(r_m + 1)$ (since we know $r_m \leq 8$).

Thus, if it is true that if $1 < |r_i - s_i| < 9$ for some positive integer i, then if $i = m$ then $r \neq s$ since $s_m \neq r_m + 1$, and if $i > m$ then $r \neq s$ since $r_i - s_i \neq 9 - 0 = 9$. Hence, $r \neq s$. □

7.4 Cardinality

Since we have access to decimals now, we can use an argument that is perhaps easier to see for the uncountability of the real numbers.

Theorem 7.16. *The open interval* $(0, 1)$ *is uncountable, as is* $\mathbb{R}$*.*

Proof. Suppose $(0, 1)$ is countable. Then there is a one to one and onto mapping $f : \mathbb{N} \to (0, 1)$ defined as follows:

$f(1) = 0.a_{11}a_{12}a_{13}...$
$f(2) = 0.a_{21}a_{22}a_{23}...$
$f(3) = 0.a_{31}a_{32}a_{33}...$

and so forth. For each $n \in \mathbb{N}$ we choose $z_n = 7$ if $a_{nn} \leq 5$ and $z_n = 3$ if $a_{nn} > 5$. Since the nth digit of the number $z = .z_1z_2z_3...$ differs from the nth digit of $f(n)$ by a number more than two and less than nine, it follows that $z \neq f(n)$ for any $n \in \mathbb{N}$ by Theorem 7.15, a contradiction to f being onto.

The real numbers contain $(0, 1)$ (and a similar argument could be performed starting with the real numbers themselves) so $\mathbb{R}$ is also uncountable.

□

Theorem 7.17. *Let* $S \subset T$*, a finite set. Then* S *is finite and* $|S| \leq |T|$*.*

Proof. Since T is finite we can find a natural number n and a one to one and onto function $f : \{1, 2, 3, ..., n\} \to T$. If S is empty then S is finite. Otherwise, let $s \in S$. If we define $F(i) = f(i)$ if $f(i) \in S$ and define $F(i) = s$ otherwise then $F : \{1, 2, 3, ..., n\} \to S$ is onto. By the well-ordering of the natural numbers there is a least natural number $m \leq n$ so that there is an onto mapping from $g : \{1, 2, 3, ..., m\} \to S$. To see that g is one to one, suppose that $g(s) = g(t)$ for some $s < t$. Then we can define a function $h : \{1, 2, 3, ..., m-1\} \to S$ by setting $h(i) = g(i)$ if $i < t$ and $h(i) = g(i+1)$ if $t \leq i \leq m-1$. For each $x \in S$ there is some $1 \leq w \leq m$ so that $g(w) = x$, so if $w < t$ then $h(w) = x$ and if $w > t$ then $h(w-1) = x$, and if $w = t$ then $h(s) = x$. Hence, h is onto, contradicting the choice of m. It follows that g is one to one and onto, so $|S| = m \leq t$.

□

Theorem 7.18. *Let* S *be a nonempty set. Then* S *is finite if and only if there is a positive integer* n *and a function* $f : \{1, 2, 3, ..., n\} \to S$ *which is onto.*

Proof. If S is finite then such a function (which is both one to one and onto) exists by definition. Assume there is a function $f : \{1, 2, 3, ..., n\} \to S$ which onto. Then there is a first integer m so that there is an onto function $g : \{1, 2, .., m\} \to S$. Suppose g is not one to one. Then there are integers $i < j$ so that $g(i) = g(j)$, in which case we can define a function $G : \{1, 2, ..., m-1\} \to S$ which is onto by setting $G(k) = g(k)$ if $k < j$ and setting $G(k) = g(k+1)$ if $k \geq j$. This contradicts our choice of m, so it follows that g is one to one and onto, and therefore $|S| = m$, which means that S is finite. □

Theorem 7.19. *A non-empty set S is countable if and only if there is an onto function $f : \mathbb{N} \to S$.*

Proof. We know that if S is countable then either S is countably infinite (in which case there is a one to one and onto function $f : \mathbb{N} \to S$ by definition), or S is finite, in which case there is a function $g : \{1, 2, 3, ..., n\} \to S$ which is onto (for some natural number n), and so if we then define $f(i) = g(1)$ for $i > n$ and $f(i) = g(i)$ for natural numbers $i \leq n$ then we have a function $f : \mathbb{N} \to S$ which is onto.

Next, assume there is a function $f : \mathbb{N} \to S$ be onto. First, assume there is an integer n so that $f(\{1, 2, ..., n\}) = S$. In this case, we will show that S is finite. By well ordering, there is a first integer m so that there is an onto function $g : \{1, 2, .., m\} \to S$. Suppose g is not one to one. Then there are integers $i < j$ so that $g(i) = g(j)$, in which case we can define a function $G : \{1, 2, ..., m - 1\} \to S$ which is onto by setting $G(k) = g(k)$ if $k < j$ and setting $G(k) = g(k + 1)$ if $k \geq j$. This contradicts our choice of m, so it follows that g is one to one and onto, and therefore $|S| = m$.

Next, assume that there is no integer n so that $f(\{1, 2, ..., n\}) = S$. In this case, we will show that S is countably infinite. Let $h(1) = f(1)$. If $h(i)$ has been chosen for all $i \leq k$ then pick $h(k+1) = f(z)$, where z is the first integer such that $f(z) \notin \{h(1), h(2), h(3), ...h(k)\}$. Note that such a z will always exist since otherwise setting $t = \max\{n \in \mathbb{N} | f(n) = h(i)$ for some $i \leq k\}$, it would follow that $f(\{1, 2, ..., t\}) = S$. Thus, $h : \mathbb{N} \to S$. We know that h is one to one because each choice of $h(j)$ was chosen to be different from $h(i)$ for $i < j$. We know that h is onto because given any $x \in S$ there is an s so that $f(s) = x$ and therefore either $h(s) = x$ or $x \in \{h(1), h(2), ..., h(s - 1)\}$ by construction. Hence, S is countably infinite.

□

Theorem 7.20. *If $|A| = n$ and $|A| = m$ then $n = m$.*

Proof. We proceed by induction on n. If $n = 1$ then there is a one to one and onto function $h : \{1\} \to A$. Thus, the definition of a function implies $A = \{h(1)\}$. Let $m > 1$ and let $r : \{1, 2, ..., m\} \to A$. Since each $r(i) = h(1)$ (the only element of A) it must follow that r is not one to one. Hence, it is false that $|A| = m$.

Assume the statement of the theorem is true whenever $n \leq k \in \mathbb{N}$. Let $f : \{1, 2, ..., k + 1\} \to A$ be a one to one and onto function, and let $g : \{1, 2, ..., m\} \to A$ be a one to one and onto function. Then defining $F(i) = f(i)$ for all $i \leq k$ defines a one to one and onto function $F : \{1, 2, ..., k\} \to A \setminus \{f(k + 1)\}$. For some j we know that $g(j) = f(k + 1)$, so we define a function $G : \{1, 2, ..., m - 1\} \to A \setminus \{f(k + 1)\}$ by setting $G(i) = g(i)$ if $i \neq j$ and setting $G(j) = g(m)$ if $j \neq m$. If $j = m$ then we just set $G(i) = g(i)$ for all $1 \leq i \leq m - 1$. Then F and G are one to one and onto functions and by the induction hypothesis, it follows that $m - 1 = k$, and therefore $m = k + 1$. By induction, the result follows.

□

Theorem 7.21. *Let A and B be sets. Then there is an onto function $f : A \to B$ if and only if there is a one to one function $g : B \to A$.*

Proof. Assume there is an onto function $f : A \to B$. For each $b \in B$ we can choose a point $a_b \in f^{-1}(b)$, and set $g(b) = a_b$. Then if $a_b = a_c$ it must follow that $b = c$ because f is a function. Hence, g is one to one.

Next, assume there is a one to one function $g : B \to A$. For each $a \in ran(g)$, define $f(a)$ to be the point b such that $g(b) = a$ (since g is one to one there is only one such choice of b). Choose an element $w \in B$. If $a \in A \setminus ran(g)$ then define $f(a) = w$. Then f is onto since $f(ran(g)) = B$.

□

Theorem 7.22. $\mathbb{N} \times \mathbb{N}$ *is countably infinite.*

Proof. Define $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ by setting $f(i, j) = 2^i 3^j$ for all $i, j \in \mathbb{N}$. By the Fundamental Theorem of Arithmetic, if $i \neq s$ or $j \neq t$ for $s, t \in \mathbb{N}$ then $2^i 3^j \neq 2^s 3^t$, which means that f is one to one, and therefore by Theorem 7.21 and Theorem 7.19 it follows that $\mathbb{N} \times \mathbb{N}$ is countable. □

Theorem 7.23. *For each $n \in \mathbb{N}$, let A_n be a countable set. Then $S = \bigcup_{n=1}^{\infty} A_n$ is countable. In other words, the union of a countable collection of countable sets is countable.*

Proof. If each A_n is empty then the result follows. Assume that at least A_1 is non-empty. For each natural number i, if $A_i \neq \emptyset$ then by Theorem 7.19, we can define an onto function $g_i : \mathbb{N} \to A_i$. By Theorem 7.22, we can find a function $h : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ which is one to one and onto. We then define $f : \mathbb{N} \times \mathbb{N} \to S$ by setting $f(i, j) = g_i(j)$ if $A_i \neq \emptyset$ and define $f(i, j) = g_1(1)$ otherwise. Then f is onto since for each $x \in S$ we know that $x \in A_i$ for some $i \in \mathbb{N}$, so $g_i(j) = x$ for some $j \in \mathbb{N}$, which means that $f(i, j) = x$. Hence, $f \circ h : \mathbb{N} \to S$ is a composition of onto functions which is onto, so S is countable by Theorem 7.19.

□

Notice that there is no requirement that the sets A_i be non-empty in the preceding proof, so in the case where only finitely many of the A_i sets are non-empty the union is just the union of the finitely many sets A_i which are non-empty. Hence, the finite union of countable sets is also countable.

We have already proven the following theorem, but the proof involved a process of choosing that might not be comfortable for some readers. Here is another proof that the rational numbers are countable using the theorems in this section.

Theorem 7.24. *The set $\mathbb{Q}$ of rational numbers is countable.*

Proof. For each $n \in \mathbb{N}$, we let $S_n = \{\frac{m}{n} \in \mathbb{Q} | m \in \mathbb{Z}\}$. Each S_n is countable. One way to observe this is to note that since $\mathbb{Z}$ is countable there is a function $g : \mathbb{N} \to \mathbb{Z}$ which is one to one and onto, so defining $G_n(i) = \frac{g(i)}{n}$ we have $G_n : \mathbb{N} \to S_n$ which is one to one and onto, so S_n is countable. Since $\mathbb{Q} = \bigcup_{n=1}^{\infty} S_n$, it follows from Theorem 7.23 that $\mathbb{Q}$ is countable.

□

Theorem 7.25. *Let S, T be disjoint finite sets. Then $|S \cup T| = |S| + |T|$. Furthermore, if $\{S_1, S_2, ..., S_m\}$ is a pairwise disjoint set of finite sets with $|S_i| = n_i$ for all $1 \leq i \leq m$ then $|\bigcup\limits_{i=1}^{m} S_i| = \sum\limits_{i=1}^{m} n_i$.*

Proof. Let $|S| = n$ and $|T| = m$. Define $h : \{1, 2, 3, ..., n\} \to \{m+1, m+2, ..., m+n\} = [m+1, m+n] \cap \mathbb{N}$ by $h(i) = i + m$. Then h is one to one since if $i < j \leq n$ then $h(i) = i + n < j + n = h(j)$, and h is onto since if j is a positive integer in $[m+1, m+n]$ then $j - n \leq m$ which means $j - n + n = m$ is in the range of h.

Since $|S| = n$ we can find $g_1 : \{1, 2, 3, ..., n\} \to S$ so that g_1 is one to one and onto. Since $|T| = m$ we can also find $g_2 : \{1, 2, 3, ..., m\} \to T$ so that g_2 is one to one and onto. We then define $f : S \cup T \to \{1, 2, 3, ..., m+n\}$ by setting $f(x) = g_1(x)$ if $x \in T$, and $f(x) = h(g_2(x))$ if $x \in S$. Then f is a well defined function since $S \cap T \neq \emptyset$. We know that f is one to one since if $x \neq y$ in $S \cup T$ then if $x \in S$ and $y \in T$ then $f(x) > m$ and $f(y) \leq m$. If $x, y \in S$ then $g_1(x) \neq g_1(y)$ since g_1 is one to one, which means that $f(x) = h(g_1(x)) \neq h(g_1(y)) = f(y)$ since h is one to one. If $x, y \in T$ then $f(x) = g_2(x) \neq g_2(y) = f(y)$ since g_2 is one to one.

Furthermore, f is onto since for every positive integer $i \leq m$ there is some $x \in T$ so that $g_2(x) = i = f(x)$ since g_2 is onto, and for every positive integer $m + 1 \leq i \leq m + n$ there is some $x \in S$ so that $g_1(x) = i - m$ and hence $f(x) = i - m + m = i$ since g_1 is onto.

Since f is one to one and onto, it follows that $|S \cup T| = |S| + |T|$.

We extend this to a finite pairwise disjoint collection $\{S_1, S_2, ..., S_m\}$ of finite sets with $|S_i| = n_i$ for all $1 \leq i \leq m$ inductively. It is certainly true that if $m = 1$ then $|S_1| = n_1$. We assume that whenever there are k many pairwise disjoint finite sets $S_1, S_2, ..., S_k$ with $|S_i| = n_i$ for all $1 \leq i \leq k$ then $|\bigcup\limits_{i=1}^{k} S_i| = \sum\limits_{i=1}^{k} n_i$. We take a collection $\{S_1, S_2, ..., S_k, S_{k+1}\}$ of pairwise disjoint finite sets with $|S_i| = n_i$ for all $1 \leq i \leq k+1$ and note by the induction hypothesis that $|\bigcup\limits_{i=1}^{k} S_i| = \sum\limits_{i=1}^{k} n_i$. Hence, since $(\bigcup\limits_{i=1}^{k} S_i) \cap S_{k+1} = \emptyset$, by the earlier part of this theorem we know that $|\bigcup\limits_{i=1}^{k+1} S_i| = |\bigcup\limits_{i=1}^{k} S_i| + |S_{k+1}| = \sum\limits_{i=1}^{k+1} n_i$. The result then follows by induction. □

The following is used often in combinatorics.

Theorem 7.26. *Pigeonhole Principle. Let $\{S_1, S_2, ..., S_m\}$ be a set of finite sets with $|S_i| = n_i$ for all $1 \leq i \leq m$. Then $|\bigcup\limits_{i=1}^{m} S_i| \leq \sum\limits_{i=1}^{m} n_i$. In other words, if $|\bigcup\limits_{i=1}^{m} S_i| > \sum\limits_{i=1}^{m} n_i$ then for at least one i it is true that $|S_i| > n_i$.*

Proof. We define $T_1 = S_1$ and inductively define $T_n = S_n \setminus \bigcup\limits_{i=1}^{n-1} S_i$ for $1 < n \leq m$. Then the sets T_n are all pairwise disjoint and since each $T_n \subseteq S_n$ we know that $|T_n| \leq |S_n| = n_i$ by Theorem 7.17. Thus, $|\bigcup\limits_{i=1}^{m} S_i| = |\bigcup\limits_{i=1}^{m} T_i| = \sum\limits_{i=1}^{m} |T_i| \leq \sum\limits_{i=1}^{m} n_i$ by Theorem 7.25. □

Theorem 7.27. *Let $A \subseteq B$. If A is uncountable then B is uncountable.*

Proof. Assume A is uncountable and $a \in A$. If B is countable then there is an onto function $f : \mathbb{N} \to \mathbb{B}$ by Theorem 7.19. We can create a map $g : B \to A$ which is onto by setting $g(x) = x$ if $x \in A$ and $g(x) = a$ otherwise. Then $g \circ f : \mathbb{N} \to A$ is onto, which means that A is countable by Theorem 7.19. This is impossible, so we conclude that B is uncountable. □

Theorem 7.28. *Let $m, n \in \mathbb{N}$. Then $|\{1, 2, 3, ..., n\} \times \{1, 2, 3, ..., m\}| = nm$.*

Proof. Fix $n \in \mathbb{N}$. Then for any natural number i, the function $g_i : \{1, 2, 3, ..., n\} \times \{i\} \to \{1, 2, 3, ..., n\}$ defined by $g_i(k) = (i, k)$ for each $1 \leq k \leq n$ is one to one and onto. Since $\{1, 2, 3, ..., n\} \times \{1, 2, 3, ..., m\} = \bigcup\limits_{i=1}^{m} \times \{1, 2, 3, ..., n\} \times \{i\}$, we know from Theorem 7.25 that

$$|\{1, 2, 3, ..., n\} \times \{1, 2, 3, ..., m\}| = \sum_{i=1}^{m} n = nm.$$

□

Definition 48

We say $|A| \leq |B|$ if and only if there is an onto function from B onto A (or, equivalently, a one to one function from A into B). We say $|A| > |B|$ (or $|B| < |A|$) if $|A| \geq |B|$ and it is false that $|B| = |A|$.

The following proof uses the Axiom of Choice, which is that every set can be ordered in a linear way (satisfying Trichotomy so that exactly one of $a < b$, $b < a$ or $a = b$ is true for each a, b in the set, and also Transitivity of order sp that if $a < b$ and $b < c$ then $a < c$ for any a, b, c in the set) which is well-ordered (meaning that every non-empty subset of the set under the order given has a least element). The process described for defining the function g is known as transfinite induction, and its validity is a consequence of the well-ordering and set specification.

Theorem 7.29. *Let A, B be sets so that it is not true that $|A| \geq |B|$. Then $|A| < |B|$.*

Proof. We use the Axiom of Choice to well-order A and B, creating a linear ordering of both sets in which every non-empty subset has a least element. We generate a function $g : A \to B$ by assigning $g(a_0) = b_0$ where a_0 is the first element of A and b_0 is the first element of B. If we have defined $g(x)$ for all $x < y$ in A then we define $g(y)$ to be the first element of B which is not equal to $g(x)$ for any $x < y$. This choice is always possible since it is false that $|A| \geq |B|$, so we know that, for any $y \in A$, with the ordering imposed by the well ordering assigned for A, that g restricted to the domain $\{x \in A | x < y\}$ is not onto (since otherwise we could define $g(x) = z$ for all other points $x \in A$ and any point $z \in B$ and we would have an onto function $g : A \to B$). Thus, for each $y \in A$ there is always some $g(y) \in B$ which is not the image of any $x < y$. Hence, $g : A \to B$ is one to one, so $|B| \geq |A|$. Since it is false that $|A| = |B|$ it follows that $|A| < |B|$. □

Theorem 7.30. *Cantor's Theorem. Let S be a set and let $P(S)$ be the set of subsets of S. Then $|P(S)| > |S|$.*

Proof. Suppose there is a function $f : S \to P(S)$ which is onto. Let $A = \{x \in S | x \notin f(x)\}$. Since f is onto there must be some $y \in S$ so that $f(y) = A$. But then $y \notin A$ since if $y \in A$ then $y \notin f(y) = a$ by definition of A. On the other hand, if $y \notin A$ then $y \notin f(y)$ so $y \in A$. Since one or the other must be true if f is onto, and both lead to a contradiction, we have a contradiction. We conclude that it is false that f is onto, so $|P(S)| > |S|$. □

Theorem 7.31. *Cantor-Schroeder-Bernstein. Let A, B be sets so that $|A| \geq |B|$ and $|B| \geq |A|$. Then $|A| = |B|$.*

Proof. Step 1. Let $D \subseteq A$ and assume there is a one to one function $h : A \to D$. Then $|A| = |D|$.

To see this, let $C_1 = A \setminus D$ and then inductively define $C_{n+1} = h(C_n)$ for each positive integer n. We wish to find $\phi : A \to D$ which is one to one and onto.

We define $\phi(x) = h(x)$ if $x \in \bigcup_{i=1}^{\infty} C_i = C$ and define $\phi(x) = x$ otherwise. Notice that the range of h is contained in D and $A \setminus D = C_1 \subseteq C$, which means that the identity map on $A \setminus C$ also has range contained in D, so the range of ϕ is contained in D.

To see that ϕ is one to one, let a, b be distinct elements of A. If $a, b \in C$ then since h is one to one $\phi(a) = h(a) \neq h(b) = \phi(b)$. If $a, b \in A \setminus C$ then $\phi(a) = a \neq b = \phi(b)$. If $a \in C$ and $b \in A \setminus C$ then $\phi(b) = b$ and for some positive integer j we know that $a \in C_j$. Thus, $\phi(a) = h(a) \in C_{j+1} \subseteq C$, which means that $\phi(a) \neq \phi(b)$. Thus, ϕ is one to one.

To see that ϕ is onto, let $d \in D$. If $d \notin C$ then $\phi(d) = d$. If $d \in C$ then $d \in C_j$ for some positive integer $j > 1$ (since d cannot be an element of C_1), which means that $d = h(s)$ for some $s \in C_{j-1}$. It follows that $|A| = |D|$.

Step 2: Let $f : A \to B$ and $g : B \to A$ be one to one functions. Then $g \circ f$ is a one to one function mapping A to $g(f(A)) \subseteq g(B)$ which means that $|A| = |g(B)|$ by Step 1. Hence, we can find a one to one and onto mapping $\psi : A \to g(B)$, so the composition $g^{-1} \circ \psi : A \to B$ is one to one and onto, and hence $|A| = |B|$.

□

7.5 More on Infinite and Sequence Limits

Many theorems regarding infinite limits have a large number of cases, depending on whether the domain is approaching infinity or a number or negative infinity, or the range is approaching infinity or negative infinity or a number. This can sometimes give rise to nine cases for some theorems, which can be cumbersome. In some theorems these cases can be consolidated by the following idea, which still requires proof in multiple cases, but this way we will need to split fewer proofs into as many cases thereafter.

Definition 49

We say that x is an *extended* real number if $x \in \mathbb{R}$ or $x = \infty$ or $x = -\infty$. Let $D \subseteq \mathbb{R}$. We will define the extended i-neighborhood about an extended real number x with respect to D, denoted $N_D(x,i)$, to be $(x - \frac{1}{i}, x + \frac{1}{i}) \cap D \setminus \{x\}$ if $x \in \mathbb{R}$, and $(i, \infty) \cap D$ if $x = \infty$ and $(-\infty, -i) \cap D$ if $x = -\infty$. If $D = \mathbb{R}$, we just write $N(x,i)$ instead of $N_{\mathbb{R}}(x,i)$. We consider the extended real number ∞ to be greater than any real number and $-\infty$ to be less than any real number. We refer to the extended real number x as being an *extended limit point* of a set D if $x \in \mathbb{R}$ and x is a limit point of D, or if $x = \infty$ and D is not bounded above or $x = -\infty$ and D is not bounded below. In other words, x is an extended real limit point of D if $N_D(x,i) \neq \emptyset$ for all $i \in \mathbb{N}$.

Note that we are not stating that ∞ or $-\infty$ exist as part of a field containing the real numbers. We are using the convention ∞ is greater than any real number and $-\infty$ is be less than any real number for reference purposes. For instance, if a theorem says $L > 0$ and L is an extended real number then L could be a positive number or the symbol L could represent ∞.

Theorem 7.32. *Let $f : D \to \mathbb{R}$. Then $\lim_{x \to c} f(x) = L$, where c, L are extended real numbers, if and only if c is an extended real limit point of D and for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c,j)$ then $f(x) \in N(L,k)$.*

Proof. Case where where $c \in \mathbb{R}$ and $L \in \mathbb{R}$: In this case, $\lim_{x \to c} f(x) = L$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < |x-c| < \delta$ then $|f(x)-L| < \epsilon$. Thus, in particular, for any $\frac{1}{k} > 0$ there is a $\delta > 0$ so that if $0 < |x-c| < \delta$ then $|f(x) - L| < \frac{1}{k}$. Choose $j \in \mathbb{N}$ so that $\frac{1}{j} < \delta$. Then if $x \in N_D(c,j)$ it follows that $f(x) \in N(L,k)$.

Conversely, assume that for any for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c,j)$ then $f(x) \in N(L,k)$. Let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that $\frac{1}{k} < \epsilon$. Then choose j so that if $x \in N_D(c,j)$ then $f(x) \in N(L,k)$. This means that if $0 < |x-c| < \frac{1}{j}$ and $x \in D$ then $|f(x) - L| < \frac{1}{k} < \epsilon$, so $\lim_{x \to c} f(x) = L$.

Case where $c \in \mathbb{R}$ and $L = \infty$ (or $-\infty$ respectively): In this case, $\lim_{x \to c} f(x) = L$ if, for every M, there is a $\delta > 0$ so that if $0 < |x-c| < \delta$ and $x \in D$ then $f(x) > M$ (respectively $f(x) < M$). In particular, for any $k \in \mathbb{N}$ we can find a $\delta > 0$ so that if $0 < |x-c| < \delta$

and $x \in D$ then $f(x) > k$ (respectively $f(x) < -k$). Choose $j \in \mathbb{N}$ so that $\frac{1}{j} < \delta$. Then if $x \in N_D(c, j)$ it follows that $f(x) \in N(L, k)$.

Conversely, for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. Let $M \in \mathbb{R}$ and choose $k \in \mathbb{N}$ so $k > m$ (respectively $-k < M$). Choose j so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. If $L = \infty$ this means that if $0 < |x-c| < \frac{1}{j}$ and $x \in D$ then $f(x) > k > M$, so $\lim_{x \to c} f(x) = \infty$. If $L = -\infty$ this means that if $0 < |x - c| < \frac{1}{j}$ and $x \in D$ then $f(x) < -k < M$, so $\lim_{x \to c} f(x) = -\infty$.

Case where $c = \infty$ (or $-\infty$ respectively) and $L \in \mathbb{R}$: In this case, $\lim_{x \to c} f(x) = L$ if and only if for every $\epsilon > 0$ there is an $M \in \mathbb{R}$ so that if $x > M$ (or $x < M$ respectively) then $|f(x) - L| < \epsilon$. Thus, in particular, for any $\frac{1}{k}$, where $k \in \mathbb{N}$, there is an M so that if $x > M$ (or $x < M$ respectively) then $|f(x) - L| < \frac{1}{k}$. Choose $j \in \mathbb{N}$ so that $j > M$ (or $-j < M$ respectively). Then if $x \in N_D(c, j)$ it follows that $f(x) \in N(L, k)$.

Conversely, assume that for any for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. Let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that $\frac{1}{k} < \epsilon$. Then choose j so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. If $c = \infty$ this means that if $x > j$ and $x \in D$ then $|f(x) - L| < \frac{1}{k} < \epsilon$, so $\lim_{x \to c} f(x) = L$. If $c = -\infty$ this means that if $x < -j$ and $x \in D$ then $|f(x) - L| < \frac{1}{k} < \epsilon$, so $\lim_{x \to c} f(x) = L$.

Case where $c = \infty$ and $L = \infty$ (or $-\infty$ respectively): In this case, $\lim_{x \to c} f(x) = L$ if, for every $M \in \mathbb{R}$, there is a B so that if $x > B$ and $x \in D$ then $f(x) > M$ (respectively $f(x) < M$). In particular, for any $k \in \mathbb{N}$ we can find a B so that if $x > B$ and $x \in D$ then $f(x) > k$ (respectively $f(x) < -k$). Choose $j \in \mathbb{N}$ so that $j > B$. Then if $x \in N_D(c, j)$ it follows that $f(x) \in N(L, k)$.

Conversely, for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. Let $M \in \mathbb{R}$ and choose $k \in \mathbb{N}$ so $k > M$ (respectively $-k < M$). Choose j so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. If $L = \infty$ this means that if $x > j$ and $x \in D$ then $f(x) > k > M$, so $\lim_{x \to c} f(x) = \infty$. If $L = -\infty$ this means that if $x > j$ and $x \in D$ then $f(x) < -k < M$, so $\lim_{x \to c} f(x) = -\infty$.

Case where $c = -\infty$ and $L = \infty$ (or $-\infty$ respectively): In this case, $\lim_{x \to c} f(x) = L$ if, for every $M \in \mathbb{R}$, there is a B so that if $x < B$ and $x \in D$ then $f(x) > M$ (respectively $f(x) < M$). In particular, for any $k \in \mathbb{N}$ we can find a B so that if $x < B$ and $x \in D$ then $f(x) > k$ (respectively $f(x) < -k$). Choose $j \in \mathbb{N}$ so that $-j < B$. Then if $x \in N_D(c, j)$ it follows that $f(x) \in N(L, k)$.

Conversely, for every $k \in \mathbb{N}$ it is true that there is some $j \in \mathbb{N}$ so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. Let $M \in \mathbb{R}$ and choose $k \in \mathbb{N}$ so $k > M$ (respectively $-k < M$). Choose j so that if $x \in N_D(c, j)$ then $f(x) \in N(L, k)$. If $L = \infty$ this means that if $x < -j$ and $x \in D$ then $f(x) > k > M$, so $\lim_{x \to c} f(x) = \infty$. If $L = -\infty$ this means that if $x < -j$ and $x \in D$ then $f(x) < -k < M$, so $\lim_{x \to c} f(x) = -\infty$.

□

Theorem 7.33. *Sequential Characterization of Limits for Extended Real Numbers (SCLE). Let c, L be extended real numbers so that c is an extended real limit point of D. Let $f : D \to \mathbb{R}$. Let $t \in \mathbb{N}$. Then $\lim\limits_{x\to c} f(x) = L$ if and only if for every sequence $\{x_n\} \subseteq N_D(c,t)$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$.*

Proof. First, assume that $\lim\limits_{x\to c} f(x) = L$. Then for every natural number k there is a natural number j so that if $x \in N_D(c,j)$ then $f(x) \in N(L,k)$ by Theorem 7.32. Let $\{x_n\} \subseteq D$ with $\{x_n\} \to c$. Then for some $k \in \mathbb{N}$, if $n \geq k$ then $x_n \in N_D(c,j)$, which means that $f(x_n) \in N(L,k)$ and hence $\{f(x_n)\} \to L$.

Next, assume that for every sequence $\{x_n\} \subseteq D$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$. Suppose $\lim\limits_{x\to c} f(x) \neq L$. Then, by Theorem 7.32, there is some $k \in \mathbb{N}$ so that for every $j \in \mathbb{N}$ it is true that there is some $x \in N_D(c,j)$ so that $f(x) \notin N(L,k)$. For each $n \in \mathbb{N}$ we choose $x_n \in N_D(c,n)$ so that $f(x_n) \notin N(L,k)$. Then $\{x_n\} \to c$, but $\{f(x_n)\} \not\to L$, which is a contradiction. We conclude that $\lim\limits_{x\to c} f(x) = L$.

□

Theorem 7.34. *Let $f, g : D \to \mathbb{R}$. Let c be an extended real number. If $\lim\limits_{x\to c} f(x) = s \in \mathbb{R}$ and $\lim\limits_{x\to c} g(x) = t \in \mathbb{R}$ then for any $\alpha, \beta \in \mathbb{R}$ it is true that $\lim\limits_{x\to c} \alpha f(x) + \beta g(x) = \alpha s + \beta t$, $\lim\limits_{x\to c} f(x)g(x) = st$, and if $t \neq 0$ and $N_{dom(\frac{f}{g})}(L,i)$ is infinite for all $i \in \mathbb{N}$ then $\lim\limits_{x\to\infty} \dfrac{f(x)}{g(x)} = \dfrac{s}{t}$.*

Proof. Let $\{x_n\}$ be a sequence of points in D so that $\{x_n\} \to c$. Then by the SCLE, we know that $\{f(x_n)\} \to s$ and $\{g(x_n)\} \to t$. Thus, $\{\alpha f(x_n) + \beta g(x_n)\} \to \alpha s + \beta t$, $\{f(x_n)g(x_n)\} \to st$ and $\{\dfrac{f(x_n)}{g(x_n)}\} \to \dfrac{s}{t}$ if $t \neq 0$ and $\{x_n\} \subset dom(\dfrac{f}{g})$. Hence, by the SCLE, we know that $\lim\limits_{x\to c} \alpha f(x) + \beta g(x) = \alpha s + \beta t$, $\lim\limits_{x\to c} f(x)g(x) = st$, and $\lim\limits_{x\to c} \dfrac{f(x)}{g(x)} = \dfrac{s}{t}$ if $t \neq 0$.

□

Theorem 7.35. *Let $f, g : D \to \mathbb{R}$, and let $\lim\limits_{x\to c} f(x) = \infty$ (respectively, $-\infty$), where c is an extended real number. Let $g(N_D(c,t))$ be bounded below (respectively, bounded above) for some natural number t. Then $\lim\limits_{x\to c} f(x) + g(x) = \infty$ (respectively, $-\infty$).*

Proof. Let $\{x_n\} \subseteq D$ with $\{x_n\} \to c$. Then by SCLE, $\{f(x_n)\} \to \infty$ (respectively $-\infty$). Choose B so that $g(N_D(c,t))$ is bounded below by B (respectively above by B). Pick $s \in \mathbb{N}$ so that if $n \geq s$ then $x_n \in N_D(c,t)$, so $g(x_n) \geq B$ (respectively $g(x_n) \leq B$). By Theorem 3.28, we know that $\{f(x_n) + g(x_n)\} \to \infty$ (respectively, $-\infty$). Thus, by SCLE we know that $\lim\limits_{x\to c} f(x) + g(x) = \infty$ (respectively, $-\infty$).

□

Theorem 7.36. *Let $f, g : D \to \mathbb{R}$.*

Let c be an extended real number and let $\lim_{x\to c} f(x) = \infty$ (respectively $-\infty$) and $\lim_{x\to c} g(x) = L \in (0,\infty)$ or $\lim_{x\to c} g(x) = \infty$ then $\lim_{x\to c} f(x)g(x) = \infty$ (respectively, $-\infty$) and $\lim_{x\to\infty} -g(x)f(x) = -\infty$ (respectively, ∞).

In particular, $\lim_{x\to c} f(x) = \infty$ if and only if $\lim_{x\to c} -f(x) = -\infty$

Proof. Let $M > 0$. Choose $k \in \mathbb{N}$ so that $k > \frac{2M}{L}$. If $\lim_{x\to c} g(x) = L \in (0,\infty)$ then also choose k so $\frac{1}{k} < \frac{L}{2}$. If $\lim_{x\to c} g(x) = \infty$ then choose k so that $k > \frac{L}{2}$. Then we can find $j \in \mathbb{N}$ so that if $x \in N_D(c,j)$ then $f(x) \in N(\infty,k)$ (respectively, $f(x) \in N(-\infty,k)$) and $g(x) \in N(\infty,k)$ if $\lim_{x\to c} g(x) = \infty$ and $g(x) \in N(L,k)$ if $\lim_{x\to c} g(x) = L$. Hence, $f(x) > \frac{2M}{L}$ (respectively $f(x) < -\frac{2M}{L}$) and $g(x) > \frac{L}{2}$, so $f(x)g(x) > M$ and $-f(x)g(x) < -M$ (respectively $f(x)g(x) < -M$ and $-f(x)g(x) > M$), which means $\lim_{x\to c} f(x)g(x) = \infty$ and $\lim_{x\to c} -f(x)g(x) = -\infty$ (respectively $\lim_{x\to c} f(x)g(x) = -\infty$ and $\lim_{x\to c} -f(x)g(x) = \infty$).

By the preceding argument, setting $g(x) = 1$, we see that if $\lim_{x\to c} f(x) = \infty$ then $\lim_{x\to c} -f(x) = -\infty$, and if $\lim_{x\to c} f(x) = -\infty$ then $\lim_{x\to c} -f(x) = \infty$.

□

Theorem 7.37. *Let c be an extended real number and let $f, g : D \to \mathbb{R}$, where f is bounded on some $N_D(c,t)$ so that for some $M > 0$ it is true that $|f(x)| < M$ is $x \in N_D(c,t)$, and let $\lim_{x\to c} g(x) = \pm\infty$. Then $\lim_{x\to c} \frac{f(x)}{g(x)} = 0$.*

Proof. Let $\{x_n\} \subseteq D$ and let $\{x_n\} \to c$. By SCLE we know that $\{g(x_n)\} \to \pm\infty$ and since for sufficiently large n we know that $x_n \in N_D(c,t)$ we know that $\{f(x_n)\}$ is bounded by Theorem 3.27, which means that $\frac{f(x_n)}{g(x_n)} \to 0$ by Theorem 3.29. Thus, by SCLE it follows that $\lim_{x\to c} \frac{f(x)}{g(x)} = 0$. □

Theorem 7.38. *Squeeze Theorem for Extended Real Numbers. Let $f, g, h : D \to \mathbb{R}$. Let $\lim_{x\to c} f(x) = L$, and let $\lim_{x\to c} h(x) = L$, where c and L are extended real numbers. If, for some natural number t is is true that $f(x) \le g(x) \le h(x)$ for all $x \in N_D(c,t)$ then $\lim_{x\to c} g(x) = L$.*

Proof. We can find some $j > t$ so that if $x \in N_D(c,j)$ then $f(x), h(x) \in N(L,k)$. Since $f(x) \le g(x) \le h(x)$ (and each $N(L,k)$ is an interval) we know that $g(x) \in N(L,k)$, which means that $\lim_{x\to c} g(x) = L$ by Theorem 7.32.

□

Definition 50

We use the convention that if $a > 0$ and $M = \infty$ and $N = -\infty$ then $aM = \infty$ and $aN = -\infty$. If $b < 0$ then we say $bM = -\infty$ and $bN = \infty$. If $c \in \mathbb{R}$ then $c + M = \infty$ and $c + N = -\infty$. We also define $MM = \infty, MN = -\infty, NN = \infty$.

Recall that $\infty, -\infty$ are not real numbers, and we don't have a multiplication operation involving them as part of a field containing the real numbers. The above notation is simply notation to make proofs and statements briefer. So, the symbols $(2)(\infty)$ in a theorem statement would simply be read as ∞.

Theorem 7.39. *Let $f, g : D \to \mathbb{R}$, where $\lim_{x\to a} f(x) = M$ and $\lim_{x\to a} f(x)g(x) = ML$, where M is a non-zero real number and a is an extended real number and L is an extended real number. Then $\lim_{x\to a} g(x) = L$. Furthermore, if $\lim_{x\to a} f(x) + g(x) = M + L$, where $M \in \mathbb{R}$, then $\lim_{x\to a} g(x) = L$.*

Proof. Let $\{x_n\} \to a$. By SCLE we know that $\{f(x_n)\} \to M \in \mathbb{R}$ and $\{f(x_n) + g(x_n)\} \to M + L$ and $\{f(x_n)g(x_n)\} \to ML$.

If $M+L \in \mathbb{R}$ then we know that $L \in \mathbb{R}$ and that $\{g(x_n)\} \to L$ by Exercise 3.8. Likewise, if $\{f(x_n)g(x_n)\} \to ML \in \mathbb{R}$ and $M \neq 0$ then $\{g(x_n)\} \to L$ by Exercise 3.9. Thus, by SCLE we know that the result is true in all cases where $M + L \in \mathbb{R}$ or $ML \in \mathbb{R}$.

Let $M + L = \infty$ (respectively $-\infty$) where $M \in \mathbb{R}$. This is true if and only if $L = \infty$ (respectively $-\infty$). Since $\{-f(x_n)\}$ converges, we know it is bounded. Thus, by Theorem 7.35 we know that $\{f(x_n) + g(x_n) - f(x_n)\} = \{g(x_n)\} \to L$, so $\lim_{x\to a} g(x) = L$ by SCLE.

Let $ML = \infty$ and $M > 0$. Then $L = \infty$. Let $k \in \mathbb{N}$. Since $M > 0$ we can find $j \in \mathbb{N}$ so that if $n \geq j$ then $\frac{M}{2} < f(x) < \frac{3M}{2}$, $f(x)g(x) > \frac{3Mk}{2}$, so $\frac{1}{f(x)} f(x_n)g(x_n) = g(x) > \frac{2}{3M}\frac{3Mk}{2} = k$, which means that $g(x) \in N(\infty, k)$, so $\lim_{x\to a} g(x) = L$.

Next, let $ML = -\infty$ and $M > 0$. Then $L = -\infty$. By Theorem 7.36, we know that $\lim_{x\to a} f(x)g(x) = -\infty$ if and only if $\lim_{x\to a} f(x)(-g(x)) = \infty$, which by the preceding argument is true if and only if $\lim_{x\to a} -g(x) = \infty$, which is true if and only if $\lim_{x\to a} g(x) = -\infty$.

Similarly, if we let $ML = \infty$ and $M < 0$ then we know that $L = -\infty$ and $\lim_{x\to a} -f(x) = -M > 0$, so we know that $\lim_{x\to a}(-f(x))g(x) = -\infty$, which means $\lim_{x\to a} g(x) = -\infty$ by the preceding argument.

Finally, if we let $ML = -\infty$ and $M < 0$ then we know that $L = \infty$ and $\lim_{x\to a} -f(x) = -M > 0$, so we $\lim_{x\to a}(-f(x))g(x) = \infty$, which means $\lim_{x\to a} g(x) = \infty$.

Thus, in all possible cases the theorem result is true.

□

Theorem 7.40. *Let c be an extended limit point of dom$(f \circ g)$. Let $\lim_{x\to c} g(x) = L$, where c and L are extended real numbers. If L is real and $f(x)$ is continuous at L then $\lim_{x\to c} f(g(x)) =$*

$f(L) = M$, a real number. If $L = \pm\infty$ and $\lim_{x \to L} f(x) = M$, an extended real number, then $\lim_{x \to c} f(g(x)) = M$.

If L is an extended limit point of the domain of f then $\lim_{x \to c} f(g(x)) = M = \lim_{y \to L} f(y)$.

Proof. This theorem is already proven for the case when c and L are real in Theorem 4.11.

Let $\{x_n\} \subseteq dom(f \circ g) \setminus \{c\}$ so that $\{x_n\} \to c$. Then by SCLE we know that $\{g(x_n)\} \to L$. If f is continuous at $L \in \mathbb{R}$ we know that $\{f(g(x_n))\} \to M$ by The Sequential Characterization of Continuity. If $L = \pm\infty$ and $\lim_{x \to L} f(x) = M$ then $\{f(g(x_n))\} \to M$ by SCLE. Thus, by SCLE we know that $\lim_{x \to c} f(g(x)) = M$.

If L is an extended limit point of the domain of $f(x)$ then by Theorem 4.4, $\lim_{y \to L} f(x) = M = \lim_{x \to c} f(g(x))$.

□

Theorem 7.41. *Let $\{a_n\}$ be a non-zero sequence so that $a_n > 0$ for sufficiently large n and $\{a_n\} \to 0$. Let $\{b_n\}$ be a non-zero sequence so that $b_n < 0$ for sufficiently large n and $\{b_n\} \to 0$. Then $\{\frac{1}{a_n}\} \to \infty$ and $\{\frac{1}{n_n}\} \to -\infty$.*

Proof. Let $M > 0$. Then we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $-\frac{1}{M} < b_n < 0 < a_n < \frac{1}{M}$, which means that $b_n < -M$ and $a_n > M$ and hence $\{\frac{1}{a_n}\} \to \infty$ and $\{\frac{1}{n_n}\} \to -\infty$. □

Theorem 7.42. *Let $\lim_{x \to c} f(x) = 0$ for a non-zero function f with domain D. Then $\lim_{x \to c} f(x) = \infty$ if f is positive in some $N_D(c,t)$ for some $t \in \mathbb{N}$, and $\lim_{x \to c} f(x) = -\infty$ if f is negative in some $N_D(c,t)$ for some $t \in \mathbb{N}$.*

Proof. Let $\{x_n\} \subseteq D$ so that $\{x_n\} \to c$. Then we know that $\{f(x_n)\} \to 0$ by the Sequential Characterization of Limits for Extended Real Numbers. If f is positive in some $N_D(c,t)$ then $f(x_n)$ is non-zero for all n and is positive for sufficiently large n, so by Theorem 7.41 we know that $\{\frac{1}{f(x_n)}\} \to \infty$, which means $\lim_{x \to c} f(x) = \infty$ by the Sequential Characterization of Limits for Extended Real Numbers. Likewise, if f is negative in some $N_D(c,t)$ then $f(x_n)$ is non-zero for all n and is negative for sufficiently large n, so by Theorem 7.41 we know that $\{\frac{1}{f(x_n)}\} \to -\infty$, which means $\lim_{x \to c} f(x) = -\infty$ □

7.6 Topology of the Real Line

We have already discussed open and closed sets and limit points. This section discusses notions of compactness and connectedness, giving ways to formulate alternate proofs of the

Extreme Value Theorem and the proof that a continuous function on a compact domain is uniformly continuous.

Definition 51

The set of open sets in a space (typically the real numbers of a subset of the real numbers for this text) is called the *topology* of the space. We say that a set $\mathcal{C}$ of sets is a *cover* of a set E if $E \subseteq \bigcup \mathcal{C}$. We say that $\mathcal{C}$ is an *open cover* of E if each element of $\mathcal{C}$ is an open set. We say that a set K is *compact* if for every open cover $\mathcal{C}$ of K there is a finite subset F of $\mathcal{C}$ which is also a cover of K. We refer to F as a finite subcover (and we may say finite subcover of K or $\mathcal{C}$, both meaning a finite subset of $\mathcal{C}$ that is a cover for K).

Theorem 7.43. *Every closed interval* $[a, b]$ *is compact.*

Proof. Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in J}$ be an open cover of $[a, b]$ (where J is an arbitrary indexing set). Let $S = \{x \in [a, b] | [a, x]$ is covered by a finite subset of $\mathcal{U}\}$. Then $a \in S$ since a is in a member of $\mathcal{U}$ and S is bounded above, so S has a least upper bound $l \in [a, b]$. Hence, for some $\beta \in J$, we know that $l \in U_\beta$, and thus for some $\epsilon > 0$ it follows that $(l - \epsilon, l + \epsilon) \subset U_\beta$ since U_β is open. By the Approximation Property, we can find some $s \in S$ so that $s > l - \epsilon$, and a finite set $F \subseteq \mathcal{U}$ so that F covers $[a, s]$ since $s \in S$. Thus, $F \cup \{U_\beta\}$ is a finite subset of $\mathcal{U}$ covering $[a, l + \epsilon)$. Since l is an upper bound for S, we know that no point of $(l, l + \epsilon)$ is in S, which implies that $l = b$ and hence $F \cup \{U_\beta\}$ is a finite subset of $\mathcal{U}$ covering $[a, b]$. Thus, $[a, b]$ is compact.

□

Theorem 7.44. *Heine-Borel Theorem. Let* $K \subset \mathbb{R}$. *Then* K *is compact if and only if* K *is closed and bounded.*

Proof. First, assume K is compact. Let $U_n = (-n, n)$ for each $n \in \mathbb{N}$. Then $\{U_n\}_{n \in \mathbb{N}}$ is an open cover for K which has a finite subcover $F = \{U_{n_1}, U_{n_2}, ..., U_{n_m}\}$, where $n_1 < n_2 < ... < n_m$. Thus, $K \subseteq \bigcup_{i=1}^{m} U_{n_i} = (-n_m, n_m)$ so K is bounded.

Suppose K is not closed. Then there is a point $p \in \overline{K} \setminus K$. Let $V_n = \mathbb{R} \setminus [p - \frac{1}{n}, p + \frac{1}{n}]$ for each $n \in \mathbb{N}$. Then $\{V_n\}_{n \in \mathbb{N}}$ is an open cover for K since $p \notin K$, and this cover has a finite subcover $G = \{V_{n_1}, V_{n_2}, ..., V_{n_j}\}$, where $n_1 < n_2 < ... < n_j$. But $\bigcup_{i=1}^{m} V_{n_i} = \mathbb{R} \setminus [p - \frac{1}{n_j}, p + \frac{1}{n_j}]$. This is impossible because $(p - \frac{1}{n_j}, p + \frac{1}{n_j}) \cap K \neq \emptyset$ since p is a limit point of K. We conclude that K must be closed.

Finally, assume that K is closed and bounded. Then for some $a < b$ we know $K \subset [a, b]$. Let $\mathcal{C} = \{C_\alpha\}_{\alpha \in J}$ be an open cover of K. Then $\mathcal{C} \cup \{\mathbb{R} \setminus K\}$ covers $[a, b]$ so that cover has a finite subcover F. Since no point of K is contained in $\mathbb{R} \setminus K$, it follows that $F \setminus \{\mathbb{R} \setminus K\}$ is a finite subset of $\mathcal{C}$ which covers K and therefore K is compact. □

Theorem 7.45. *Let $f : E \to \mathbb{R}$. Then:*

(a) The function f is continuous if and only if, for every open subset V of $\mathbb{R}$, the set $f^{-1}(V)$ is open in E.

(b) The function f is continuous at a point $p \in E$ if and only if for every open set V containing $f(p)$, there is an open set U containing p such that $U \cap E \subseteq f^{-1}(V)$.

Proof. (a) First, assume that f is continuous. Let V be open in $\mathbb{R}$ and let $x \in f^{-1}(V)$. Since V is open there is an $\epsilon > 0$ so that $(x - \epsilon, x + \epsilon) \subseteq V$. Since f is continuous there is a $\delta > 0$ so that if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$ for all $\mathbf{y} \in E$. Thus $(x - \delta, x + \delta) \cap E \subseteq f^{-1}(V)$, so by $f^{-1}(V)$ is open in E.

Next, assume that the inverse of every open set is open in E. Let $p \in E$ and let $\epsilon > 0$. Note that $(p - \epsilon, p + \epsilon)$ is open, so $f^{-1}((p - \epsilon, p + \epsilon))$ is open in E, which means that there is a $\delta > 0$ so that $(p - \delta, p + \delta) \cap E \subseteq f^{-1}((p - \epsilon, p + \epsilon))$. Hence, if $|x - p| < \delta$ and $x \in E$ then $|f(x) - f(p)| < \epsilon$, which means that f is continuous.

(b) First, assume that f is continuous at $p \in E$. Let V be open in $\mathbb{R}$ containing $f(p)$. Since V is open there is an $\epsilon > 0$ so that $(x - \epsilon, x + \epsilon) \subseteq V$. Since f is continuous there is a $\delta > 0$ so that if $|x - p| < \delta$ then $|f(x) - f(p)| < \epsilon$ for all $x \in E$. Thus $(p - \delta, p + \delta) \cap E \subseteq f^{-1}(V)$.

Next, assume that for every open set V containing $f(p)$, there is an open set U containing p such that $U \cap E \subseteq f^{-1}(V)$. Let $\epsilon > 0$. Note that $(f(p) - \epsilon, f(p) + \epsilon)$ is open. Thus, there is an open set U so that $U \cap E \subset f^{-1}((f(p) - \epsilon, f(p) + \epsilon))$, which means that there is a $\delta > 0$ so that $(p - \delta, p + \delta) \cap E \subseteq U \cap E \subseteq f^{-1}((f(p) - \epsilon, f(p) + \epsilon))$. Hence, if $|x - p| < \delta$ and $x \in E$ then $|f(x) - f(p)| < \epsilon$, which means that f is continuous at p. □

Theorem 7.46. *Let $f : K \to \mathbb{R}$ be continuous, where K is compact. Then $f(K)$ is compact.*

Proof. Let $C = \{U_\alpha\}_{\alpha \in J}$ be an open cover of $f(K)$. For each $\alpha \in J$, since $f^{-1}(U_\alpha)$ is open in K we may choose V_α open in $\mathbb{R}$ so that $V_\alpha \cap K = f^{-1}(U_\alpha)$. Then the $\{V_\alpha\}_{\alpha \in J}$ sets covers K and has a finite subcover $V_{\alpha_1}, V_{\alpha_2}, .., V_{\alpha_t}$. The corresponding open sets $U_{\alpha_1}, U_{\alpha_2}, .., U_{\alpha_t}$ are a finite subset of C which covers $f(K)$. Thus $f(K)$ is compact. □

Definition 52

We say that a continuous one to one and onto function $f : E \to K$ is a *homeomorphism* if $f^{-1} : K \to E$ is also continuous, in which case we say that E and K are *homeomorphic* (topologically indistinguishable, essentially).

Theorem 7.47. *Let $f : K \to \mathbb{R}$ be continuous and one to one, where K is compact. Then $f^{-1} : f(K) \to \mathbb{R}$ is also continuous. In other words, f is a homeomorphism.*

Proof. Let A be a closed set. Then $A \cap K$ is closed since A and K are closed, and is bounded since K is bounded, so A is compact by the Heine-Borel Theorem. By Theorem 7.46, we know that $f(A \cap K)$ is compact and therefore closed. Since the inverse image of A under f^{-1} is $f(A \cap K)$ it follows that f^{-1} is continuous on $f(K)$ by Exercise 7.7. □

Theorem 7.48. *The Extreme Value Theorem. Let $f : K \to \mathbb{R}$ be continuous, where K is closed and bounded. Then there are points $s, t \in K$ so $f(s) \leq f(x) \leq f(t)$ for every $x \in K$.*

Proof. By the Heine-Borel Theorem, K is compact, so we know that $f(K)$ is compact by Theorem 7.46. By Theorem 3.22 this means that $f(K)$ has a largest value $f(t)$ and a smallest value $f(s)$ for some $s, t \in K$. □

Definition 53

Let L be the set of limit points of a set E. The *closure* of a set E, denoted $\overline{E}$ is $E \cup L$. A pair of non-empty subsets A and B of a set E is a *separation* of E if $A \cup B = E$ and the sets $A \cap \overline{B} = \emptyset = \overline{A} \cap B$. A set E is *connected* if it has no separation. If I is a bounded interval then we use the notation $|I|$ to refer to the *length* of I (the right end point of I minus the left end point of I). We say a set $D \subseteq S$ is *dense* in S if $\overline{D} \supseteq S$. We say that S is *separable* if S has a countable subset which is dense in S.

Theorem 7.49. *Let $f : E \to \mathbb{R}$ be continuous and let E be connected. Then $f(E)$ is connected.*

Proof. Suppose $f(E)$ has a separation A, B. Then since A, B are disjoint, $f^{-1}(A)$ and $f^{-1}(B)$ are disjoint. Since $\overline{A} \cap B = \emptyset$ and $\overline{A}$ is closed by Exercise 4.5, we know that $f^{-1}(B) = f^{-1}(\mathbb{R} \setminus \overline{A}$ is open in E. Likewise, $f^{-1}(A)$ is open in E. Since A, B are non-empty, $f^{-1}(A), f^{-1}(B)$ are non-empty. Furthermore, if $p \in f^{-1}(B)$ and a sequence $\{x_n\} \subseteq A$ and $\{a_n\} \to p$, it follows that $\{f(x_n\} \to f(p)$ which means that $f(p)$ is a limit point of A which is contained in B, which is impossible. Hence, $f^{-1}(B)$ contains no limit points of $f^{-1}(A)$ and likewise $f^{-1}(A)$ contains no limit points of $f^{-1}(B)$. Thus, the pair $f^{-1}(A), f^{-1}(B)$ is a separation of E, which is impossible since E is connected. □

Theorem 7.50. *Let A, B be non-empty disjoint subsets of a set E so that $A \cup B = E$. Then the pair of sets A, B is a separation of E if and only if there are disjoint open sets U, V so that $U \cap E = A$ and $V \cap E = B$, which is true if and only if A is both closed and open in E.*

Proof. First, assume that A, B form a separation of E. Then no point of B is a limit point of A and no point of B is a limit point of A, which means that for each $x \in A$ there is some $\epsilon_x > 0$ so that $(x - \epsilon_x, x + \epsilon_x) \cap B = \emptyset$ and for each point $y \in B$ there is some $\epsilon_y > 0$ so that $(y - \epsilon_y, y + \epsilon_y) \cap A = \emptyset$. Let $U = \bigcup_{x \in A}(x - \frac{\epsilon_x}{3}, x + \frac{\epsilon_x}{3})$ and let $V = \bigcup_{y \in B}(y - \frac{\epsilon_y}{3}, y + \frac{\epsilon_y}{3})$. Then $A \subseteq U$ and $B \subseteq V$. Suppose that $z \in U \cap V$ then for some $x_0 \in A$ and $y_0 \in B$, $|z - x_0| < \frac{\epsilon_{x_0}}{3}$ and $|z - y_0| < \frac{\epsilon_{y_0}}{3}$. Assume that $\epsilon_{x_0} \geq \epsilon_{y_0}$. Then $|x_0 - y_0| \leq |x_0 - z| + |z - y_0| < \frac{\epsilon_{x_0}}{3} + \frac{\epsilon_{y_0}}{3} \leq \frac{2\epsilon_{x_0}}{3}$, which is impossible since $(x_0 - \epsilon_{x_0}, x_0 + \epsilon_{x_0}) \cap B = \emptyset$. If $\epsilon_{x_0} \leq \epsilon_{y_0}$ we arrive at a similar contradiction. We conclude that $U \cap V = \emptyset$.

Assume that there are disjoint open sets U, V so that $U \cap E = A$ and $V \cap E = B$. Then A and B are both open in E. Since $B = E \setminus A = (\mathbb{R} \setminus U) \cap B$ and $A = E \setminus B = (\mathbb{R} \setminus V) \cap E$ we know that A, B are both closed in E.

Assume that A is open and closed in E. Then since A is closed in E, there is a closed set K so that $K \cap E = A$, which means that $(\mathbb{R} \setminus K) \cap E = B$ is open in E. Since A is open in E there is an open set U so that $U \cap E = A$ which means that each point of A is contained in an open interval that does not intersect B, so no point of A is a limit point of B and $A \cap \overline{B} = \emptyset$. Similarly, $B \cap \overline{A} = \emptyset$ since B is open in E. Hence, A and B form a separation of E.

□

Theorem 7.51. *Let $E \subseteq \mathbb{R}$. Then E is connected if and only if E is an interval.*

Proof. First, assume E is not an interval. Then for some $a, b \in E$ it follows that $a < c < b$ and $c \notin E$ for some point c. But then $(-\infty, c) \cap E, (c, \infty) \cap E$ are a separation of E, so E is not connected.

Next, assume that E is an interval and suppose that E is not connected. Then E has a separation A, B. Define $f : E \to \mathbb{R}$ by $f(x) = 0$ if $x \in A$ and $f(x) = 2$ if $x \in B$. Since for the inverse image of every set under f is one of the empty set, A, B or all of E, all of which are open in E, we know that f is continuous. By the Intermediate Value Theorem, it follows that $f(c) = 1$ for some $c \in (a, b)$, which is impossible. We conclude that E is connected. □

Theorem 7.52. *Lebesgue Number Lemma. Let K be a compact set and $\mathcal{C} = \{C_\alpha\}_{\alpha \in J}$ be an open cover of K. Then there is a number $\delta > 0$ so that if I is an open interval such that $I \cap K \neq \emptyset$ and $|x - y| < \delta$ for all $x, y \in I$ then $I \subseteq C_\beta$ for some $\beta \in J$.*

Proof. Since $\mathcal{C}$ is an open cover of K, for each $p \in K$ we may choose ϵ_p so that $(p - 2\epsilon_p, p + 2\epsilon_p) \subseteq C_{\alpha_p}$ for some $\alpha_p \in J$. Then $\{(p - \epsilon_p, p + \epsilon_p)\}_{p \in K}$ is an open cover of K with finite subcover $F = \{(p_i - \epsilon_{p_i}, p + \epsilon_{p_i})\}_{1 \leq i \leq m}$. Let $\delta = \min\{\epsilon_{p_i}\}_{1 \leq i \leq m}$. Let I be an interval with $|I| < \delta$ and let $x \in I \cap K$. Then for some j we know that $x \in (p_j - \epsilon_{p_j}, p_j + \epsilon_{p_j})$ so if $y \in I$ then $|p_j - y| \leq |p_j - x| + |x - y| < \epsilon_{p_j} + \delta < 2\epsilon_{p_j}$. Thus, $I \subseteq (p_j - 2\epsilon_{p_j}, p_j + 2\epsilon_{p_j}) \subseteq C_{\alpha_{p_j}}$.
□

Note that, in particular if x, y are points of K with $|x-y|<\delta$ then since $[x, y]$ is a subset of an element of the cover $\mathcal{C}$ it follows that both x and y are in one element of $\mathcal{C}$.

Theorem 7.53. *Let $f : K \to R$ be continuous, where K is closed and bounded. Then f is uniformly continuous.*

Proof. Let $\epsilon > 0$. For each $x \in K$ choose $\delta_x > 0$ so that if $y \in K$ and $|x-y| < \delta_x$ then $|f(x)-f(y)| < \frac{\epsilon}{2}$. Note that $\mathcal{C} = \{(x-\delta_x, x+\delta_x)|x \in K\}$ is an open cover of K. Since K is compact (by the Heine Borel Theorem) we know that there is a number $\delta > 0$ so that if $x, y \in K$ and $|x-y| < \delta$ then for some $z \in K$ it follows that $x, y \in (z-\delta_z, z+\delta_z)$, so $|f(x)-f(y)| \le |f(x)-f(z)| + |f(z)-f(y)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. □

7.7 L'Hospital's Rule

We will be using the notation developed in the section "More on Infinite Limits" in the following argument.

Theorem 7.54. *L'Hospital's Rule. Let $f, g : I \to \mathbb{R}$ be differentiable on an interval I and let $g'(x) \neq 0$ for all $x \in I \setminus \{a\}$ where a is either an element of the open interval I or an end point of I or $a = \infty$ and I is not bounded above or $a = -\infty$ and I is not bounded below. Let $\lim_{x\to a} f(x) = 0 = \lim_{x\to a} g(x)$ and $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$ or let $\lim_{x\to a} f(x) = \pm\infty$ and let $\lim_{x\to a} g(x) = \pm\infty$ and $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$, where L is an extended real number. Then $\lim_{x\to a} \frac{f(x)}{g(x)} = L$.*

Proof. First, we observe that we proved this theorem for the cases where $L \in \mathbb{R}$ and $\lim_{x\to a} f(x) = 0 = \lim_{x\to a} g(x)$ in chapter 5.

Next, assume that $\lim_{x\to a} f(x) = \pm\infty$ and let $\lim_{x\to a} g(x) = \pm\infty$ and $\lim_{x\to a} \frac{f'(x)}{g'(x)} = L$, where a and L are extended real numbers.

Choose $k_0 \in \mathbb{N}$ so that if $x \in N_I(a, k_0)$ then $|g(x)| > 0$. Choose a point $s_1 \in N_I(a, k_0)$. Since $\lim_{x\to a} g(x) = \pm\infty$ we can find $k_1 > k_0$ so that if $t \in N_I(a, k_1)$ then $\frac{|g(s_1)|}{|g(t)|} < 1$ and $\frac{|f(s_1)|}{|g(t)|} < 1$. Let $\{t_n\} \subset N_I(a, k_1)$, where $\{t_i\} \to a$. We then find $k_2 > k_1$ so that if $t \in N_I(a, k_2)$ then $\frac{|g(t_1)|}{|g(t)|} < \frac{1}{2}$ and $\frac{|f(t_1)|}{|g(t)|} < \frac{1}{2}$. For some integer m_1 it is true that if $n \ge m_1$ then $t_n \in N_I(a, k_2)$. We set $s_i = s_1$ if $1 \le i < m_1$. Note that $\frac{|g(s_i)|}{|g(t_i)|} < 1$ and $\frac{|f(s_i)|}{|g(t_i)|} < 1$ if $1 \le i < m_1$. We then find $k_3 > k_2$ so that if $t \in N_I(a, k_3)$ then $\frac{|g(t_2)|}{|g(t)|} < \frac{1}{3}$ and $\frac{|f(t_2)|}{|g(t)|} < \frac{1}{3}$. We then find $m_2 > m_1$ so that if $n \ge m_2$ then $t_n \in N_I(a, k_3)$. We define $s_i = t_1$

if $m_1 \le i < m_2$. Note that $\dfrac{|g(s_i)|}{|g(t_i)|} < \dfrac{1}{2}$ and $\dfrac{|f(s_i)|}{|g(t_i)|} < \dfrac{1}{2}$ if $m_1 \le i < m_2$ (since $g(s_i) = g(t_1)$ and $f(s_i) = f(t_1)$ if $m_1 \le i < m_2$). We then find $k_4 > k_3$ so that if $t \in N_I(a, k_4)$ then $\dfrac{|g(t_3)|}{|g(t)|} < \dfrac{1}{4}$ and $\dfrac{|f(t_3)|}{|g(t)|} < \dfrac{1}{4}$. We then find $m_3 > m_2$ so that if $n \ge m_3$ then $t_n \in N_I(a, k_4)$. We define $s_i = t_2$ if $m_2 \le i < m_3$. Note that $\dfrac{|g(s_i)|}{|g(t_i)|} < \dfrac{1}{3}$ and $\dfrac{|f(s_i)|}{|g(t_i)|} < \dfrac{1}{3}$ if $m_2 \le i < m_3$.

We continue in this manner, creating a sequence of points $\{s_i\}$ with a corresponding increasing sequence of natural numbers $\{m_j\}$ so that $s_i = t_j$ for all $m_j \le i < m_{j+1}$, with the m_j integers chosen so that whenever $i \ge m_j$ it is true that $\dfrac{|g(s_i)|}{|g(t_i)|} < \dfrac{1}{j}$ and $\dfrac{|f(s_i)|}{|g(t_i)|} < \dfrac{1}{j}$. Then $\{s_i\} \to a$ (since for $i > m_j$ we know $s_i \in N(a, k_j) \subseteq N(a, j)$), $\dfrac{|g(s_i)|}{|g(t_i)|} \to 0$ and $\dfrac{|f(s_i)|}{|g(t_i)|} \to 0$ (since for $i \ge m_j$ we know that $0 \le \dfrac{|g(s_i)|}{|g(t_i)|} < \dfrac{1}{j}$ and $0 \le \dfrac{|f(s_i)|}{|g(t_i)|} < \dfrac{1}{j}$).

By the Cauchy Mean Value Theorem, for each $n \in \mathbb{N}$ we can choose c_n between s_n and t_n so that $f'(c_n)(g(t_n) - g(s_n)) = g'(c_n)(f(t_n) - f(s_n))$. Then $\dfrac{f'(c_n)}{g'(c_n)} = \dfrac{f(t_n) - f(s_n)}{g(t_n) - g(s_n)} = \dfrac{\frac{f(t_n)}{g(t_n)} - \frac{f(s_n)}{g(t_n)}}{\frac{g(t_n)}{g(t_n)} - \frac{g(s_n)}{g(t_n)}} = (\dfrac{f(t_n)}{g(t_n)} - \dfrac{f(s_n)}{g(t_n)})(\dfrac{1}{1 - \frac{g(s_n)}{g(t_n)}})$. Since $\{\dfrac{f'(c_n)}{g'(c_n)}\} \to L$, and $\{\dfrac{1}{1 - \frac{g(s_n)}{g(t_n)}}\} \to 1$, by Exercise 3.9, we know that $\{\dfrac{f(t_n)}{g(t_n)} - \dfrac{f(s_n)}{g(t_n)}\} \to L$. Since $\{\dfrac{f(s_n)}{g(t_n)}\} \to 0$, by Theorem 7.39 we know that $\{\dfrac{f(t_n)}{g(t_n)}\} \to L$. Hence, by SCLE we know that $\lim\limits_{x \to a} \dfrac{f'(x)}{g'(x)} = L$.

□

Theorem 7.55. *L'Hospital's rule can be extended to sequences. Let* $f, g : [1, \infty) \to \mathbb{R}$ *be differentiable functions, where* g *is non-zero. Let* $\{f(n)\}, \{g(n)\}$ *be sequences defined by restricting* f *and* g *to the natural numbers. Assume* $\lim\limits_{n \to \infty} f(x_n) = 0 = \lim\limits_{n \to \infty} f(x_n)$ *or* $\lim\limits_{n \to \infty} f(x_n) = \pm\infty$ *and* $\lim\limits_{n \to \infty} g(x_n) = \pm\infty$. *Also assume that* $\lim\limits_{n \to \infty} \dfrac{f'(x)}{g'(x)} = L$. *Then it follows that* $\lim\limits_{n \to \infty} \dfrac{f(x_n)}{g(x_n)} = L$.

Proof. By L'Hospital's rule, we know that $\lim\limits_{x \to \infty} \dfrac{f(x)}{g(x)} = L$. Thus, by the Sequential Characterization of Limits for Extended Real Numbers, we have $\lim\limits_{n \to \infty} \dfrac{f(x_n)}{g(x_n)} = L$.

□

Usually, when we use L'Hospital's rule for sequences we above we tend to abuse notation by not observing that the sequences can be extended to functions on intervals and then using this theorem to conclude that the limit of the ratio of the sequences is the same as the limit of the ratio of the functions. Instead, we tend to treat n as a variable representing any number in $[1, \infty)$, use L'Hospital's rule and then treat the limit as the limit of the sequence itself afterwards. It should be understood that this is what we are doing, however, since otherwise it doesn't make much sense to use L'Hospital's rule (since a function defined only on the integers is not differentiable at all). We don't normally quote this theorem, and instead just use it without mentioning it.

7.8 More on Integration

Change of Variables for Single Variable Functions:

Theorem 7.56. *Let $\phi : U \to \mathbb{R}$ be C^1 on $[a, b]$, where U is an open set containing $[a, b]$ with $\phi'(x) \neq 0$ on $[a, b]$.*

(a) There is an interval $[c, d]$ so that $\phi([a, b]) = [c, d]$.

(b) ϕ^{-1} has continuous non-zero derivative on $[c, d]$.

(c) There is an $M > 0$ so that for any two points $x, y \in [a, b]$, $|\phi(x) - \phi(y)| < M|x - y|$ and for any two points $x, y \in [c, d]$, $|\phi^{-1}(x) - \phi^{-1}(y)| < M|x - y|$.

(d) If $D \subset [a, b]$ and $\lambda(D) = 0$ then $\lambda(\phi(D)) = 0$. If $E \subset [c, d]$ and $\lambda(E) = 0$ then $\lambda(\phi^{-1}([c, d])) = 0$.

(e) Let f be integrable on $[c, d]$. Then $f \circ \phi$ is integrable on $[a, b]$.

Proof. (a) First, $\phi(E)$ is a closed interval $[c, d]$ by Exercise 4.13.

(b) By Theorem 5.26, we know that ϕ is strictly monotone and by the Inverse Function Theorem we know that ϕ^{-1} is differentiable with a non-zero derivative $\dfrac{1}{\phi'(\phi^{-1}(x))}$ on $[c, d]$. Since $\phi'(x)$ is continuous and positive, and $\phi^{-1}(x)$ is continuous, it follows that the derivative of $\phi^{-1}(x)$ is continuous on $[c, d]$.

(c) Let $\psi = \phi^{-1}$. Since ϕ', ψ' are non-zero and continuous on $[a, b]$, by the Extreme Value Theorem these functions are bounded, so we can find M so that $M > |\phi'(x)|$ for all $x \in [a, b]$ and $M > |\psi'(x)|$ for all $x \in [c, d]$. Hence, by Exercise 5.10, we know that for any $x, y \in [a, b]$,if $x \neq y$ then $|\phi(x) - \phi(y)| < M|x - y|$ and if $x, y \in [c, d]$ then $|\phi^{-1}(x) - \phi^{-1}(y)| < M|x - y|$.

(d) By (c) we can find $M > 0$ so that for any two points $x, y \in [a, b]$, $|\phi(x) - \phi(y)| < M|x - y|$ and for any two points $x, y \in [c, d]$, $|\phi^{-1}(x) - \phi^{-1}(y)| < M|x - y|$.

Assume $\lambda(D) = 0$ for some $D \subseteq [a, b]$. Let $\epsilon > 0$ and choose a collection of closed intervals $\{I_1\}_{i\in\mathbb{N}}$ covering D so that $\displaystyle\sum_{i=1}^{\infty} |I_i| < \frac{\epsilon}{M}$. By (a) we know that each $\phi(I_i)$ is a closed interval and for any two points x, y in I_i we know that $|\phi(x) - \phi(y)| < M|x - y|$, which means that $\phi(I_i)$ is a closed interval whose end points are closer together than $M|I_i|$ and therefore $|\phi(I_i)| < M|I_i|$. Hence, $\{\phi(I_i)\}_{i\in\mathbb{N}}$ is a cover of $\phi(D)$ and $\displaystyle\sum_{i=1}^{\infty} |\phi(I_i)| < M \sum_{i=1}^{\infty} |I_i| < M\frac{\epsilon}{M} = \epsilon$. Thus, $\lambda(\phi(D)) = 0$. Since ϕ^{-1} is also a C^1 one to one function by (b) it follows that if $\lambda(E) = 0$ for some $E \subseteq [c, d]$ then $\phi^{-1}(E)$ also has Lebesgue measure zero.

(e) Let D_f be the set of points of $[c, d]$ at which f is not continuous. We know that $\lambda(D_f) = 0$ by the Lebesgue Characterization of Riemann Integrability, which means that $\lambda(\phi^{-1}(D_f)) = 0$ by part (d). If $x \in [a, b] \setminus \phi^{-1}(D_f)$ then ϕ is continuous at x and f is continuous at $f(x)$, which means that the set $D_{f\circ\phi}$ of all discontinuities of $f \circ \phi$ in $[a, b]$ is a subset of $\phi^{-1}(D_f)$ of measure zero, from which we conclude that $f \circ \phi$ is integrable on $[a, b]$.

□

The following is a stronger form of the substitution result for compositions of integrable functions with continuously differentiable functions.

Theorem 7.57. *Change of variables, single variable case.*

Let ϕ be continuously differentiable on $E = [a, b]$ with $\phi'(x) \neq 0$ on $[a, b]$. Let f be an integrable function on $\phi([a, b])$. Then $\int_E f \circ \phi|\phi'| = \int_{\phi(E)} f$.

Proof. First, note that $f \circ \phi$ is integrable on E, and we can find $M > 1$ so that for any two points $x, y \in [a, b]$, $|\phi(x) - \phi(y)| < M|x - y|$ and for any two points $x, y \in [c, d]$, $|\phi^{-1}(x) - \phi^{-1}(y)| < M|x - y|$ by Theorem 7.56. This also means that $f \circ \phi|\phi'|$ is integrable on $[a, b]$ by Theorem 6.11.

By Theorem 6.6 we can find a $\delta > 0$ so that if P is a partition of $[a, b]$ with $|P| < \delta$ then $U(f \circ \phi|\phi'|, P) - L(f \circ \phi|\phi'|, P) < \frac{\epsilon}{2}$, and if Q is a partition of $[c, d]$ so that $|Q| < \delta$ then $U(f, Q) - L(f, Q) < \frac{\epsilon}{2}$.

By 5.26, we know that ϕ is strictly monotone. For now, we will assume that ϕ is increasing and so $|\phi'(x)| = \phi'(x) > 0$ on $[a, b]$ since it is impossible for $\phi'(x) < 0$ for any x for an increasing function by Exercise 5.12.

Choose a partition $P = \{p_0, p_1, p_2, ..., p_n\}$ of $[a, b]$ so that $|P| < \frac{\delta}{M}$. Let $Q = \{q_0, q_1, q_2, ..., q_n\}$ be a partition of $[c, d]$ where $q_i = \phi(p_i)$ for all $0 \leq i \leq n$. Since $|q_i - q_{i-1}| \leq M|p_i - p_{i-1}| < M\frac{\delta}{M} = \delta$, we know $U(f \circ \phi|\phi'|, P) - L(f \circ \phi|\phi'|, P) < \epsilon$, and $U(f, Q) - L(f, Q) < \epsilon$. Thus, any Riemann sum of $f \circ \phi|\phi'|$ with respect to partition P has distance less than $\frac{\epsilon}{2}$ from $\int_a^b f \circ \phi|\phi'|$, and any Riemann sum of f with respect to partition Q has distance less than $\frac{\epsilon}{2}$ from $\int_c^d f$.

By the Mean Value Theorem we can find a marking $R = \{p_i^*\}_{1 \leq i \leq n}$ so that $p_i^* \in (p_{i-1}, p_i)$ and $\phi'(p_i^*)(p_i - p_{i-1}) = \phi(p_i) - \phi(p_{i-1}) = q_i - q_{i-1}$ for all $1 \leq i \leq n$. Let $T = \{\phi(p_i^*)\}_{1 \leq i \leq n}$, which is a marking of Q since ϕ is increasing. Then $S_T(f, Q) = \sum_{i=1}^{n} f(\phi(p_i^*))(q_i - q_{i-1}) = \sum_{i=1}^{n} f(\phi(p_i^*)\phi'(p_i^*)(p_i - p_{i-1}) = S_R(f \circ \phi|\phi'|, P)$.

It follows that $|\int_c^d f - \int_a^b f \circ \phi|\phi'|| \leq |\int_c^d f - S_T(f, Q)| + |S_T(f, Q) - S_R(f \circ \phi|\phi'|, P)| + |S_R(f \circ \phi|\phi'|, P) - \int_a^b f \circ \phi|\phi'|| < \frac{\epsilon}{2} + 0 + \frac{\epsilon}{2} = \epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $\int_c^d f = \int_a^b f \circ \phi|\phi'|$.

If ϕ is decreasing the proof is similar. We have $\phi'(x) < 0$ on $[a,b]$ and we choose P and Q as before except that $q_i = \phi(p_{n-i})$. As before, we can find a marking $R = \{p_i^*\}_{1\leq i\leq n}$ so that $p_i^* \in (p_{i-1}, p_i)$ and $\phi'(p_i^*)(p_i - p_{i-1}) = \phi(p_i) - \phi(p_{i-1}) = q_{n-i} - q_{n-i+1}$ for all $1 \leq i \leq n$. Let $T = \{\phi(p_i^*)\}_{1\leq i\leq n}$, which is a marking of Q since ϕ is decreasing. Then $S_T(f,Q) = \sum_{i=1}^{n} f(\phi(p_i^*))(q_i - q_{i-1}) = -\sum_{i=1}^{n} f(\phi(p_i^*)\phi'(p_i^*)(p_i - p_{i-1}) = -S_R((f \circ \phi)\phi', P) = S_R((f \circ \phi)|\phi'|, P)$ since $|\phi'| = -\phi$.

Thus, as before, $|\int_c^d f - \int_a^b f \circ \phi|\phi'|| \leq |\int_c^d f - S_T(f,Q)| + |S_T(f,Q) - S_R(f \circ \phi|\phi'|, P)| + |S_R(f \circ \phi|\phi'|, P) - \int_a^b f \circ \phi|\phi'|| < \frac{\epsilon}{2} + 0 + \frac{\epsilon}{2} = \epsilon$, and since this is true for all $\epsilon > 0$ we conclude that $\int_c^d f = \int_a^b f \circ \phi|\phi'|$.

. □

Improper Integration:

Definite Riemann integrals are only defined for bounded functions over closed, bounded intervals. We can extend the idea of an integral to unbounded intervals or to functions whose ranges are unbounded with the idea of an improper integral, which is the limit of a definite integral or the sum of multiple such limits.

If one bound of an integral is infinity (or negative infinity respectively) then we can replace that bound by a variable, evaluate the resulting integral as normal and let the bound approach infinity (or negative infinity respectively). If the resulting limit exists then we say that the improper integral exists (converges) and is equal to the specified limit. If not then we say that the improper integral diverges (it does not exist). If the interval over which the integral is taken is from negative infinity on the lower bound to infinity on the upper bound then you must take a point in between and split the integral into two integrals each of which only approaches an infinite limit at one bound. If either integral diverges then we say the integral is divergent. If both converge then the integral is the sum of the resulting integrals.

Definition 54

Let $f : D \to \mathbb{R}$. If $[a, \infty) \subseteq D$ then we say that the *improper integral* $\int_a^\infty f(x)dx = \lim_{b\to\infty} \int_a^b f(x)dx$ if this limit exists.

If $(-\infty, a] \subseteq D$ then we say that improper integral $\int_{-\infty}^a f(x)dx = \lim_{b\to-\infty} \int_b^a f(x)dx$ if this limit exists.

If $(c, d] \subseteq D$ and f is not integrable on $[c, d]$ then we say that improper integral $\int_c^d f(x)dx = \lim_{b\to c^+} \int_b^d f(x)dx$ if this limit exists.

If $[a, c) \subseteq D$ and f is not integrable on $[a, c]$ then we say that improper integral $\int_a^c f(x)dx = \lim_{b\to c^-} \int_a^c f(x)dx$ if this limit exists.

If an improper integral integral of f exists on an interval D is equal to I then we also say that the improper integral converges to I, and write $\int_D f = I$. If the improper integral does not exist we refer to the improper integral as divergent.

If D is a finite union of intervals, no two of which intersect at more than one point, on which the improper integrals of f all exist, then the sum of the improper integrals of f on each interval is defined to be the improper integral of f on D. Otherwise, the improper integral of f on D does not exist (it is divergent).

Example 7.1. *Find the improper integral* $\int_1^\infty \frac{1}{x^3}dx$.

Solution. Our definition of improper integral is that this is $\lim_{b\to\infty} \int_1^b \frac{1}{x^3}dx = \lim_{b\to\infty} \left.\frac{-1}{2x^2}\right|_1^b =$ $\lim_{b\to\infty} \frac{1}{2} - \frac{2}{2b^2} = \frac{1}{2}$.

□

Example 7.2. *Find the improper integral* $\int_{-\infty}^\infty e^{-x}dx$.

Solution. We pick a point to separate the integral into a sum of two integrals which are improper at only one bound. We will choose $\int_{-\infty}^0 e^{-x}dx + \int_0^\infty e^{-x}dx$. Integrating we have $\int_0^\infty e^{-x}dx = \lim_{b\to\infty} \int_0^b e^{-x}dx = \lim_{b\to\infty} -e^{-x}\big|_0^b = \lim_{b\to\infty} -e^{-b} + 1 = 1$.

For the other integral we have $\lim_{b\to-\infty} \int_b^0 e^{-x}dx = \lim_{b\to-\infty} -e^{-x}\big|_b^0 = \lim_{b\to-\infty} e^{-b} - 1 =$ ∞. Thus, $\int_{-\infty}^\infty e^{-x}dx$ diverges.

□

In the case where there is a finite set of points at which the integrand approaches infinity or negative infinity, we separate the integral into a sum of integrals over subintervals of the interval over which the original integral is taken, each of which is only approaching infinity (or is infinity) at one bound of the integral and at no other point on the subinterval over which each integral defined. For each integral where one bound is a point at which the integrand approaches infinity or negative infinity, we replace that bound by b and take the limit as b approaches the original bound from the direction so that b is in the interval over which the integral was taken. If any of these new integrals diverges we say the original integral diverges. If all of them converge then the sum of the integrals on these subintervals is the value of the original integral.

It is sometimes helpful to compare two functions and determine the convergence or divergence of the integral of one function's improper integral by comparing it to the other. In order to handle all cases at once, we will use some of our development from the limits at extended real numbers from earlier in the Supplementary Materials.

Theorem 7.58. *Let D be a an interval in the domain of f and g so that either D has finite length and f and g approach $\pm\infty$ at exactly one end point of D, or D is either unbounded above or below (but not both) and f and g do not approach infinity at any point of D. Let $0 \le f(x) \le g(x)$ on D. Then if $\displaystyle\int_D g(x)dx = L$, a finite value, then $\displaystyle\int_D f(x)dx = M$, a finite value so that $M \le L$. Likewise, if $\displaystyle\int_D f(x)dx = \infty$ then $\displaystyle\int_D g(x)dx = \infty$.*

Proof. We know that $\displaystyle\int_D f = \lim_{b\to c}\int_{D(b)} f$, where c is an extended real number and $D(b)$ is an interval contained in D, one of whose ends is c, and the other is the other (finite) end point of D at which f and g do not approach infinity. Since f and g are non-negative on D we know that if b_2 is closer to c than b_1 is to c (where what is meant by "closer" $c = \infty$ for this proof is that $b_2 > b_1$ what is meant by "closer" if $c = -\infty$ is that $b_2 < b_2$) then $\displaystyle\int_{D(b_1)} f \le \int_{D(b_2)} f$ since $\displaystyle\int_{D(b_2)} f - \int_{D(b_1)} f = \int_I f \ge 0$, where I is the interval whose end points are b_1 and b_2. Likewise, $\displaystyle\int_{D(b_1)} g \le \int_{D(b_2)} g$.

Choose a sequence $\{x_n\} \subset D$ so that $\{x_n\} \to c$ and for each natural number n it is true that x_{n+1} is closer to c than x_n.

First, assume that $\displaystyle\int_D g(x)dx = L$. Then by the Sequential Characterization of Limits for Extended Real Numbers, we know that $\{\displaystyle\int_{D_{x_n}} g\} \to L$, where $\{\displaystyle\int_{D_{x_n}} g\}$ is a non-decreasing sequence whose supremum is L. The sequence $\{\displaystyle\int_{D_{x_n}} f\}$ is also non-decreasing and is bounded above by L, which means that $\{\displaystyle\int_{D_{x_n}} f\} \to M = \sup_{n\in\mathbb{N}}\displaystyle\int_{D_{x_n}} f$ for some $M \le L$. This means that given any $\epsilon > 0$ there is an x_n so that $M - \displaystyle\int_{D_{x_n}} f < \epsilon$. If b in the interior of D and is closer to c than x_n then there is an

x_j closer to c than b, which means $L - \epsilon < \int_{D_{x_n}} f \leq \int_{D_b} f \leq \int_{D_{x_j}} f \leq L$, so it follows that $\lim_{b \to c} \int_{D(b)} f = M$.

Finally, assume that $\int_D f(x)dx = \infty$. Then by the Sequential Characterization of Limits for Extended Real Numbers, we know that $\{\int_{D_{x_n}} f\} \to \infty$. By The Squeeze Theorem for Extended Real Numbers we know that $\{\int_{D_{x_n}} g\} \to \infty$. This means that given any $R > 0$ there is an x_n so that $\int_{D_{x_n}} g > R$. If b in the interior of D and is closer to c than x_n then $R < \int_{D_{x_n}} g \leq \int_{D_b} g$, so it follows that $\lim_{b \to c} \int_{D(b)} g = \infty$.

□

It is sometimes helpful in both series, and in many infinite limits or limits at infinity to note the following relative sizes of increase. We will refer to this as the "Order of Bigness List" (it has no formal name that is generally accepted, so we will use this silly name which we hope is unlikely to be duplicated as meaning something else in another context). Some aspects of the associated theorem we prove here are useful now, and others are useful after we cover series. To describe them it is helpful to introduce a new notation.

Definition 55

We say that a function $f(x)$ converges at a rate less than that of a positive constant times $g(x)$, a function which is positive for sufficiently large values of x, or that $f(x)$ is big O of $g(x)$, written $f(x)$ *is* $O(g(x))$ if there is some $k \in \mathbb{N}$ and $c > 0$ so that $|f(x)| < cg(x)$ for all $x \geq k$. This is also written $f(x) = O(g(x))$.

Note that while the "=" sign is part of an acceptable notation for the "big O" notation described above, it does not mean equality in this context. It is immediate that if f is $O(g(x))$ and $h(x) > g(x)$ for all sufficiently large x then f is $O(h(x))$.

Theorem 7.59. *The Order of Bigness List. On domain* $[1, \infty)$, *for the following (ordered) list of functions:* $\ln(x), x^p$ *where* $0 < p < 1$, x^q *where* $q > 1$, r^x *where* $r > 1$, $\lfloor x \rfloor !$, *and* x^x*:*

(a) Each function on the list approaches infinity as x *approaches infinity.*

(b) If a function f *precedes (is listed earlier in the list than) a function* g *then* $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$. *It is also true that* $\lim_{x \to \infty} \frac{1}{\ln(x)} = 0$.

(c) With the exception of the pair consisting of $ln(x)$ *and* x^p *where* $0 < p < 1$*, and the pair* x^p *where* $0 < p < 1$ *and* x^q *where* $q > 1$ *if* $q - p \leq 1$*, if* f *precedes* g *in this list then there is a number* $k > 1$ *so that* $\dfrac{f(x)}{g(x)}$ *is* $O(\dfrac{1}{x^k})$*.*

Proof. By Theorem 6.25, $\ln(x)$ is increasing and is not bounded above, which means that for any $M > 0$ there is some $k > 0$ so that if $x > k$ then $\ln(x) > M$ and therefore $\lim\limits_{x\to\infty} \ln(x) = \infty$.

Next, we look at x^p for $0 < p$. Taking the natural log, we see get $p\ln(x)$. Since $p > 0$ and $\lim\limits_{x\to\infty} \ln(x) = \infty$, by Theorem 7.36. Hence, given any $M > 0$ we can find $k > 1$ so that if $x \geq k$ then $p\ln(x) > \ln(M)$ and thus $e^{p\ln(x)} > M$, so $x^p > M$, which means that $\lim\limits_{x\to\infty} \ln(x) = \infty$.

Next, if $r > 1$ then note that $\ln(r) > 0$ by Theorem 6.25, so $(r^x)' = r^x \ln(r) >$ which means that r^x is increasing.

Let $M > 0$. If $x > \dfrac{\ln(M)}{\ln(r)}$ it follows that $r^x > r^{\frac{\ln(M)}{\ln(r)}} = e^{\frac{\ln(M)}{\ln(r)}\ln(r)} = e^{\ln(M)} = M$ by definition. Thus, $\lim\limits_{x\to\infty} r^x = \infty$.

Next, note that $\lfloor x \rfloor! \geq (n-1)!$ on for each natural number $n \geq x$. Let $M > 0$. Since $\mathbb{N}$ is not bounded above we can find $k > M + 1$, so if $x \geq k$ then $\lfloor x \rfloor! \geq (k-1)! = (k-1)(k-2)...(1) \geq M$. Thus, $\lim\limits_{x\to\infty} \lfloor x \rfloor! = \infty$.

Finally, for any $M > 0$ we can pick $k > \max\{1, M\}$. If $x \geq k$ then $x\ln(x) \geq x\ln(k)$ since $x > 0$ and $\ln(x)$ is increasing. Since e^x is increasing, this means that $e^{x\ln(x)} \leq e^{x\ln(k)}$, so $x^x \geq k^x$. We have already established that $\lim\limits_{x\to\infty} k^x = \infty$, so by the Squeeze Theorem for Extended Real Numbers we see that $\lim\limits_{x\to\infty} x^x = \infty$

(b) and (c) Since we know that $\lim\limits_{x\to\infty} \ln(x) = \infty$, it follows from Theorem 7.37 that $\lim\limits_{x\to\infty} \dfrac{1}{\ln(x)} = 0$.

Let $p > 0$. Using L'Hospital's Rule, we see that $\lim\limits_{x\to\infty} \dfrac{\ln(x)}{x^p} = \lim\limits_{x\to\infty} \dfrac{\frac{1}{x}}{px^{p-1}} = \lim\limits_{x\to\infty} \dfrac{1}{p}\dfrac{1}{x^p} = 0$ by Theorem 7.37 since we know that $\lim\limits_{x\to\infty} px^p = \infty$ from part (a) and Theorem 7.36.

Let $q > 1$ and $0 < p < 1$. Then $\lim\limits_{x\to\infty} \dfrac{x^p}{x^q} = \lim\limits_{x\to\infty} \dfrac{1}{x^{q-p}} = 0$ by Theorem 7.37 since we know that $\lim\limits_{x\to\infty} x^{q-p} = \infty$ from part (a). In the case where $q - p > 1$ then $\dfrac{x^p}{x^q}$ is $O(\dfrac{1}{x^{q-p}})$ where $q - p > 1$. Likewise, if $q > 1$ then we can pick p so that $0 < p < q - 1$, and notice that $\dfrac{\ln(x)}{x^q} = \dfrac{\ln(x)}{x^p}\dfrac{x^p}{x^q}$, and for sufficiently large x we know that $\dfrac{\ln(x)}{x^p} < 1$, so $\dfrac{\ln(x)}{x^p}$ is $O(\dfrac{1}{x^{q-p}})$ where $q - p > 1$.

Let $p > 1$ and $q > 0$. Then we can find a natural number m so that $m - 1 < q \leq m$. By L'Hospital's rule, $\lim\limits_{x\to\infty} \dfrac{x^q}{p^x} = \lim\limits_{x\to\infty} \dfrac{q(q-1)x^{q-2}}{p^x(\ln(p))^2}$, and so on, until the numerator no longer approaches infinity, so the limit is equal to $\lim_{x\to\infty} \frac{q(q-1)...(q-m+1)x^{q-m}}{p^x(\ln(p))^m} = 0$ by Theorem 7.37.

Since q was an arbitrary positive number we can replace q by $q+2$, so we can choose k so that if $x \geq k$ then $\frac{x^{q+2}}{p^x} = x^2\frac{x^q}{p^x} < 1$, which means that $\frac{x^q}{p^x} < \frac{1}{x^2}$. Thus, $\frac{x^q}{p^x}$ is $O(\frac{1}{x^2})$. Likewise, since $\frac{\ln(x)}{p^x} < \frac{x^r}{p^x} < \frac{x^q}{p^x}$ for sufficiently large x for $0 < r < 1$, it follows that $\frac{\ln(x)}{p^x}$ and $\frac{x^r}{p^x}$ are also $O(\frac{1}{x^2})$.

Next, let $p > 1$ and choose $m \in \mathbb{N}$ so $m > 2p+1$. Then if $m+k < x < m+k+1$, $\lfloor x \rfloor ! \geq (m+k)(m+k-1)...(m+1)m!$, so $\frac{p^x}{\lfloor x \rfloor !} \leq \frac{p^{k+1}}{(m+k)(m+k-1)...(m+1)}\frac{p^m}{m!} = (\frac{p}{m+k})(\frac{p}{m+k-1})...(\frac{p}{m+1})(\frac{p^{m+1}}{m!}) < (\frac{1}{2})^k(\frac{p^{m+1}}{m!})$. We know that $\{(\frac{1}{2})^k\} \to 0$ by Exercise 3.13, so for any $\epsilon > 0$ we can choose $t \in \mathbb{N}$ so that if $n \geq t$ then $(\frac{1}{2})^n < \frac{m!\epsilon}{p^{m+1}}$. Thus, if $x > t$ then $\frac{p^x}{\lfloor x \rfloor !} < (\frac{1}{2})^n(\frac{p^{m+1}}{m!}) < \epsilon$ and hence $\lim_{x\to\infty} \frac{p^x}{\lfloor x \rfloor !} = 0$. This also establishes that $\frac{p^x}{\lfloor x \rfloor !}$ is $O(\frac{1}{2}^n)$. Since we also know from the preceding paragraph that $\frac{1}{2}^n$ is $O(\frac{1}{x}^2)$ and $\ln(x) < x^r < x^q < p^x$ for sufficiently large x, this means that $\frac{\ln(x)}{\lfloor x \rfloor !}, \frac{x^r}{\lfloor x \rfloor !}$ and $\frac{x^q}{\lfloor x \rfloor !}$ are each $O(\frac{1}{x^2})$.

Finally, if $x > 2$ then $\frac{\lfloor x \rfloor !}{x^x} = \frac{(\lfloor x \rfloor)(\lfloor x \rfloor - 1)(\lfloor x \rfloor - 2)...(1)}{(x)(x)(x)...(x)(x^{x-\lfloor x \rfloor})} = (\frac{\lfloor x \rfloor}{x})(\frac{\lfloor x \rfloor - 1}{x})(\frac{\lfloor x \rfloor - 2}{x})...(\frac{2}{x})(\frac{1}{x})(\frac{1}{x^{x-\lfloor x \rfloor}}) < \frac{2}{x^2}$, so $\frac{\lfloor x \rfloor !}{x^x}$ is $O(\frac{1}{x^2})$.

Since we have shown $\ln(x) < x^r < x^q < p^x < x^x$ for $0 < r < 1$, $q > 1$ and $p > 1$ (for sufficiently large x) we conclude that each of $\frac{\ln(x)}{x^x}, \frac{x^r}{x^x}, \frac{x^q}{x^x}$ and $\frac{p^x}{x^x}$ are each $O(\frac{1}{x^2})$.

□

Lebesgue Characterization of Riemann Integrability:

Definition 56

We say that a set S has *Lebesgue measure zero*, denoted $\lambda(S) = 0$ if and only if for every $\epsilon > 0$ there is a sequence of open intervals $\{I_i\}$ which covers S so that $\sum_{i=1}^{\infty} |I_i| \leq \epsilon$ (where $|I_i|$ denotes the length of interval I_i).

Let $f : D \to \mathbb{R}$ be bounded. For each non-degenerate interval I we define the *oscillation of f on I* to be $\Omega_f I = \sup\limits_{x\in I\cap D} f(x) - \inf\limits_{x\in I\cap D} f(x)$. If $p \in \overline{E}$ then we define the *oscillation of f at p* to be $\omega_f(p) = \lim\limits_{h\to 0^+} \Omega_f(p-h, p+h)$.

Theorem 7.60. *Let $A \subset B$ and let $\lambda(B) = 0$. Then $\lambda(A) = 0$.*

Proof. Let $\epsilon > 0$. Then there is an open cover of B by a countable collection of open intervals $\{I_n\}$ so that $\sum\limits_{n=1}^{\infty} |I_n| < \epsilon$. Since $\{I_n\}$ is also a cover for A it follows that $\lambda(A) = 0$. □

Theorem 7.61. *Let $E = \{p_1, p_2, p_3, ...\}$ be countable. Then $\lambda(E) = 0$.*

Proof. Let $I_i = (p_i - \frac{\epsilon}{2^{i+2}}, p_i + \frac{\epsilon}{2^{i+2}})$. Then $\{I_i\}$ covers E and $\sum\limits_{i=1}^{\infty} |I_i| = \epsilon$, so $\lambda(E) = 0$. □

Theorem 7.62. *Let $\lambda(E_i) = 0$ for each $i \in \mathbb{N}$. Then $\lambda(\bigcup\limits_{i=1}^{\infty} E_i) = 0$.*

Proof. For each $i \in \mathbb{N}$ choose open intervals $\{I_{(i,n)}\}_{n\in\mathbb{N}}$ which cover E_i so that $\sum\limits_{n=1}^{\infty} |I_{(i,n)}| < \frac{\epsilon}{2^{i+1}}$. Then $\sum\limits_{i=1}^{\infty}\sum\limits_{j=1}^{\infty} |I_{(i,j)}| < \epsilon$ and $\{I_{(i,j)}\}_{i,j\in\mathbb{N}}$ is a cover for $\bigcup\limits_{i=1}^{\infty} E_i$, which has Lebesgue measure zero. □

Theorem 7.63. *If we remove "open" or replace "open" by "closed" in our definition of Lebesgue measure zero and the definitions would be equivalent.*

Proof. Let S be a set and let $\epsilon > 0$. First, assume $\lambda(S) = 0$. Then we can find a countable cover of S by open intervals $\{I_i\}_{i\in\mathbb{N}}$ so that $\sum\limits_{i=1}^{\infty} |I_i| < \epsilon$. If we add the end points to each I_i the sum of the lengths of the intervals is unchanged and the intervals still cover S.

Next, assume that for every $\epsilon > 0$ we can find a countable cover by closed intervals (or simply intervals) $\{I_i\}_{i\in\mathbb{N}}$ so that $\sum\limits_{i=1}^{\infty} |I_i| < \epsilon$. Then choose intervals $\{A_i\}$ so that $\sum\limits_{i=1}^{\infty} |A_i| < \frac{\epsilon}{2}$. Let the left and right end points of A_i be a_i and b_i respectively. Then define $V_i = (a_i - \frac{\epsilon}{2^{i+2}}, b_i + \frac{\epsilon}{2^{i+2}})$. Then $\{V_i\}_{i\in\mathbb{N}}$ is a countable collection of open intervals which covers S so that $\sum_{i=1}^{\infty} |V_i| < \epsilon$, so $\lambda(S) = 0$. □

Theorem 7.64. *$f : D \to \mathbb{R}$ be bounded and let I_1, I_2 be intervals with $I_1 \subseteq I_2$. Then $\Omega_f I_1 \leq \ \Omega_f I_2$.*

Proof. We know $\sup\limits_{x\in I_1\cap D} f(x) \leq \sup\limits_{x\in I_2\cap D} f(x)$ and $\inf\limits_{x\in I_1\cap D} f(x) \geq \inf\limits_{x\in I_2\cap D} f(x)$ by Exercise 1.17, which means $\Omega_f I_1 \leq \ \Omega_f I_2$. □

Theorem 7.65. *$f : D \to \mathbb{R}$ be bounded and let $p \in \overline{D}$. Then $\omega_f(p) = \inf\limits_{h\in\mathbb{R}^+} \Omega_f(p - h, p + h) \geq 0$.*

Proof. Note that $\Omega_f(p - h, p + h)$ is a non-negative real number for each $h > 0$ since $\{f(x)|x \in (p - h, p + h)\}$ is non-empty and bounded if $p \in \overline{E}$. Let $w = \inf\limits_{h\in\mathbb{R}^+} \Omega_f(p - h, p + h)$. Then $w \geq 0$ since each $\Omega_f(p - h, p + h) \geq 0$. Let $\epsilon > 0$. Then for some $\delta > 0$ we know that $\Omega_f(p - \delta, p + \delta) < w + \epsilon$ by the approximation property. However, we also know that if $0 < h < \delta$ then $\Omega_f(p - h, p + h) \leq \Omega_f(p - \delta, p + \delta)$ by Theorem 7.64, so $|\Omega_f(p-h, p+h) - w| < \epsilon$ and $w = \omega_f(p) = \lim\limits_{h\to 0^+} \Omega_f(p-h, p+h)$. □

Theorem 7.66. *Let $f : D \to \mathbb{R}$ be bounded and let $p \in D$. Then f is continuous at p if and only if $\omega_f(p) = 0$.*

Proof. Assume f is continuous at p and let $\epsilon > 0$. Choose $\delta > 0$ so that if $|x - p| < \delta$ and $x \in D$ then $|f(x) - f(p)| < \frac{\epsilon}{2}$. Then $\sup\limits_{x\in(p-\delta,p+\delta)\cap D} f(x) \leq f(p) + \frac{\epsilon}{2}$ and $\inf\limits_{x\in(p-\delta,p+\delta)\cap D} f(x) \geq f(p) - \frac{\epsilon}{2}$. Thus $\Omega_f(p - \delta, p + \delta) \leq \epsilon$ so $\omega_f(p) \leq \epsilon$ for all $\epsilon > 0$, which means that $\omega_f(p) = 0$.

Assume that $\omega_f(p) = 0$ and let $\epsilon > 0$. Then since $\omega_f(p) = \inf\limits_{h\in\mathbb{R}^+} \Omega_f(p - h, p + h)$, by the Approximation Property we can find $\delta > 0$ so that $\Omega_f(p - \delta, p + \delta) < \epsilon$, which means that if $|x - c| < \delta$ and $x \in D$ then $|f(x) - f(c)| < \epsilon$, so f is continuous at c. □

Theorem 7.67. *Let K be a closed set, and $f : K \to \mathbb{R}$ be bounded, and let $E_n = \{x \in K|\omega_f(x) \geq \frac{1}{n}\}$. Then E_n is closed. If K is compact then E_n is compact.*

Proof. Let $\{x_n\} \subseteq E_n$, where $\{x_n\} \to p$. Then for any $h > 0$ we know that $(p - \frac{h}{2}, p + \frac{h}{2})$ contains x_m for some $m \in \mathbb{N}$. Since $\frac{1}{n} \leq \omega_f(x_m) \leq \Omega_f(x_m - \frac{h}{2}, x_m + \frac{h}{2}) \leq \Omega_f(p-h, p+h)$ by Theorems 7.65 and 7.64, it follows that $\omega_f(p) \geq \frac{1}{n}$. Hence, E_n contains all of its limit points and is closed. If K is compact then E_n is also bounded and thus compact by the Heine-Borel Theorem. □

Theorem 7.68. *Let $f : D \to \mathbb{R}$ be bounded and let $p \in \overline{D}$ and $\epsilon > 0$. If $\omega_f(p) < \epsilon$ then there is a $\delta > 0$ so that $\Omega_f(p - \delta, p + \delta) < \epsilon$.*

Proof. This follows directly from the Approximation Property since we know that $\omega_f(p) = \inf_{\{h \in \mathbb{R} | h > 0\}} \Omega_f(p - h, p + h)$. □

Theorem 7.69. *Let $f : K \to \mathbb{R}$ be a bounded function, with K a compact set. Let $\epsilon > 0$ and $\omega_f(p) < \epsilon$ for each $p \in K$. Then there is a $\delta > 0$ so that if I is an interval so that $|I| < \delta$ and $I \cap K \neq \emptyset$ then $\Omega_f I < \epsilon$.*

Proof. By theorem 7.68 for each $p \in K$ we can find a $\epsilon_p > 0$ so that $\Omega_f(p - \epsilon_p, p + \epsilon_p) < \epsilon$. Then $\mathcal{C} = \{(p - \epsilon_p, p + \epsilon_p)\}_{p \in K}$ is an open cover of K, so by the Lebesgue Number Lemma we can find $\delta > 0$ so that if I is an interval and $I \cap K \neq \emptyset$ and $|I| < \delta$ then $I \subseteq (p - \epsilon_p, p + \epsilon_p)$ for some $p \in K$, which means that $\Omega_f I < \epsilon$. □

Theorem 7.70. *Let S be a set and $\epsilon > 0$ such that for every countable $\{I_i\}_{i \in \mathbb{N}}$ of open intervals which cover S, $\sum_{i=1}^{\infty} |I_i| \geq \epsilon$. If $\{U_i\}_{i \in \mathbb{N}}$ is any countable collection of intervals which covers S then $\sum_{i=1}^{\infty} |U_i| \geq \epsilon$.*

Proof. Suppose that there are intervals $\{U_i\}_{i \in \mathbb{N}}$ which cover S so that $\sum_{i=1}^{\infty} |U_i| = r < \epsilon$, where the left and right end points of U_i are a_i and b_i respectively. Then define $V_i = (a_i - \frac{\epsilon - r}{2^{i+2}}, b_i + \frac{\epsilon - r}{2^{i+2}})$, and $\{V_i\}_{i \in \mathbb{N}}$ is a countable collection of open intervals which covers S so that $\sum_{i=1}^{\infty} |V_i| = r + \frac{\epsilon - r}{2} < \epsilon$, a contradiction.

□

Theorem 7.71. *Let E be a set and $\gamma > 0$ so that for any sequence of intervals $\{I_i\}$ which covers E, $\sum_{i=1}^{\infty} |I_i| \geq \gamma$. Let $C = \{I_i\}$ be such a cover. Then $\sum_{i \in \mathbb{N} | I_i^{\circ} \cap E \neq \emptyset} |I_i| \geq \gamma$.*

Proof. Suppose that $\sum_{\{i | I_i^0 \cap E \neq \emptyset\}} |I_i| = \alpha < \gamma$. Then all of the points of E not covered by $C = \{I_i | I_i^0 \cap E \neq \emptyset\}$ are end points a_i, b_i of the I_i intervals. Thus, if we set $B = \bigcup_{i=1}^{\infty} \{a_i, b_i\}$ then $\lambda(B) = 0$ since B i countable, and so we can find a countable collection of intervals $D = \{R_i\}_{i \in \mathbb{N}}$ which cover B so that $\sum_{i=1}^{\infty} |R_i| < \gamma - \alpha$. Hence, the set of all intervals in $C \cup D$ is a countable collection of intervals which cover E, the sum of whose volumes is less than γ, a contradiction.

□

Theorem 7.72. *Let $P = \{x_1, x_2, ..., x_s\}$ be a partition of $[a, b]$.*

Let $C = \{[x_{n_1 - 1}, x_{n_1}], [x_{n_2 - 1}, x_{n_2}], ..., [x_{n_t - 1}, x_{n_t}]\}$, where $1 \leq n_1 < n_2 < ... < n_t$ and let $K = \bigcup C$. Let $D = \{I_1, I_2, ..., I_m\}$ be a cover of K by open intervals $I_i = (a_i, b_i)$. Then $\sum_{i=1}^{m} |I_i| > \sum_{i=1}^{t} x_{n_i} - x_{n_{i-1}}$.

Proof. Since K is a finite union of closed sets K is closed, and since $K \subseteq [a, b]$, K is bounded, so K is compact by the Heine-Borel Theorem.

By the Lebesgue Number Lemma we can find $\delta > 0$ so that if I is a closed interval of length less than δ and I intersects K then I is a subset of some $I_i \in D$.

Next, we choose a partition Q of $[a, b]$ that refines P and has mesh less than δ, where $Q = \{q_0, q_1, q_2, ..., q_w\}$. Then $\sum_{i=1}^{t} x_{n_i} - x_{n_{i-1}} = \sum_{i=1}^{t} \sum_{\{j \in \mathbb{N} | [q_{j-1}, q_j] \subseteq [x_{n_{i-1}}, x_{n_i}]\}} q_j - q_{j-1}$
$\leq \sum_{i=1}^{m} \sum_{\{j \in \mathbb{N} | [q_{j-1}, q_j] \subseteq I_i\}} q_j - q_{j-1} < \sum_{i=1}^{m} |I_i|$.

The reason the last inequality is strict is that, for each i, if p is the least integer so that $[q_p, q_{p+1}] \subseteq I_i$ and r is the last integer so that $[q_r, q_{r+1}] \subseteq I_i$ then $a_i < p$ and $r < b_i$ which means $b_i - a_i > q_{r+1} - q_p = \sum_{i=p}^{r} q_{i+1} - q_i = \sum_{\{j \in \mathbb{N} | [q_{j-1}, q_j] \subseteq I_i\}} q_j - q_{j-1}$.

□

We did not list justifications for the other parts of the expression $\sum_{i=1}^{t} x_{n_i} - x_{n_{i-1}} = \sum_{i=1}^{t} \sum_{\{j \in \mathbb{N} | [q_{j-1}, q_j] \subseteq [x_{n_{i-1}}, x_{n_i}]\}} q_j - q_{j-1} \leq \sum_{i=1}^{m} \sum_{\{j \in \mathbb{N} | [q_{j-1}, q_j] \subseteq I_i\}} q_j - q_{j-1} < \sum_{i=1}^{m} |I_i|$ since they are similar to things we have observed earlier, but to clarify further, since Q refines P, for a given interval $[x_{n_{i-1}}, x_{n_i}]$ there is some k so that $q_k = x_{n_{i-1}}$ and some r so that $q_{r+1} = x_{n_i}$, which tells us that $\sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq[x_{n_{i-1}},x_{n_i}]\}} q_j - q_{j-1} = (q_{k+1} - q_k) + (q_{k+2} - q_{k+1}) + ... + (q_{r+1} - q_r) = q_{r+1} - q_k = x_{n_i} - x_{n_{i-1}}$. This justifies the first equality.

For the first inequality, we notice that each $q_{j+1} - q_j$ in the preceding sum is added once (since $[q_{j+1} - q_j]$ cannot be a subset of different intervals $[x_{n_{i-1}}, x_{n_i}]$). Each such $[q_{j+1} - q_j]$ is a subset of some I_i by the Lebesgue Number Lemma since $q_{j+1} - q_j < \delta$ and $[q_{j+1} - q_j] \cap K \neq \emptyset$. Thus, every term $q_{j+1} - q_j$ in the sum $\sum_{i=1}^{t} \sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq[x_{n_{i-1}},x_{n_i}]\}} q_j - q_{j-1}$ is a summand in the sum $\sum_{i=1}^{m} \sum_{\{j\in\mathbb{N}|[q_{j-1},q_j]\subseteq I_i\}} q_j - q_{j-1}$. It is possible that such a term may appear more than once in the second sum, since it is possible for $[q_j - q_{j-1}]$ to be a subset of more than one I_i. It is also possible that there are intervals $[q_j - q_{j-1}]$ contained in an element of D which do not appear in the left sum. Either of these two possibilities would result in a larger sum on the right. Thus, the inequality follows.

The following theorem is one of the most helpful results for deciding when a function is Riemann integrable.

Theorem 7.73. *Lebesgue Characterization of Riemann Integrability. Let $f : [a, b] \to \mathbb{R}$ be bounded. Then f is integrable if and only if the set $E = \{x \in [a, b]|f$ is not continuous at $x\}$ has Lebesgue measure zero.*

Proof. Since f is bounded we can choose $M > 0$ so that $|f(x)| < M$ for all $x \in [a, b]$. For each $n \in \mathbb{N}$ let $E_n = \{x \in [a, b]|\omega_f(x) \geq \frac{1}{n}\}$. Note that $E = \bigcup_{n=1}^{\infty} E_n$. If $\lambda(E) = 0$ then $\lambda(E_n) = 0$ for each $n \in \mathbb{N}$ by Theorem 7.60.

Assume that f is integrable. Suppose that $\lambda(E) \neq 0$. Then for some $m \in \mathbb{N}$ we know from Theorem 7.62 that $\lambda(E_m) \neq 0$, so there is a number $\gamma > 0$ so that if $\{I_i\}_{i\in\mathbb{N}}$ is an open cover of E_m then $\sum_{i=1}^{\infty} |I_i| \geq \gamma$. Let $P = \{x_0, x_2, x_3, ..., x_k\}$ be a partition of $[a, b]$. Then $U(f, P) - L(f, P) = \sum_{\{i\in\mathbb{N}|(x_{i-1},x_i)\cap E_m\neq\emptyset\}} (M_i - m_i)(x_i - x_{i-1}) + \sum_{\{i\in\mathbb{N}|(x_{i-1},x_i)\cap E_m=\emptyset\}} (M_i - m_i)(x_i - x_{i-1})$. By Theorem 7.65 we know $M_i - m_i \geq \frac{1}{m}$ if $(x_{i-1}, x_i) \cap E_m \neq \emptyset$, so we know that $\sum_{\{i\in\mathbb{N}|(x_{i-1},x_i)\cap E_m\neq\emptyset\}} (M_i - m_i)(x_i - x_{i-1}) \geq \frac{\gamma}{m}$ since $\sum_{\{i\in\mathbb{N}|(x_{i-1},x_i)\cap E_m\neq\emptyset\}} (x_i - x_{i-1}) \geq \gamma$ by Theorem 7.71, and thus f is not integrable, a contradiction.

Assume that $\lambda(E) = 0$. Let $\epsilon > 0$. Choose $j \in \mathbb{N}$ so that $\frac{b-a}{j} < \frac{\epsilon}{2}$. Choose a countable cover of E_j by open intervals $\{I_i\}_{i\in\mathbb{N}}$ so that $\sum_{i=1}^{\infty} |I_i| < \frac{\epsilon}{4M}$. Since E_j is compact by Theorem 7.67, we can find a finite subcover $F = \{I_{n_1}, I_{n_2}, ..., I_{n_t}\}$. Let $K = [a,b] \backslash \bigcup_{i=1}^{t} I_{n_i}$. We know K is compact by the Heine-Borel Theorem, and $\omega(p) < \frac{1}{j}$ for all $p \in K$, so by by Theorem 7.68 we can find a number $\delta > 0$ so that if I is an interval intersecting K with $|I| < \delta$ then $\Omega_f I < \frac{1}{j}$.

Let $P = \{x_1 x_2, ..., x_s\}$ be a partition of $[a,b]$ with $|P| < \delta$. Then $U(f,P) - L(f,P) = \sum_{\{i\in\mathbb{N}|[x_{i-1},x_i]\cap K \neq \emptyset\}} (M_i - m_i)(x_i - x_{i-1}) + \sum_{\{i\in\mathbb{N}|[x_{i-1},x_i]\cap K = \emptyset\}} (M_i - m_i)(x_i - x_{i-1})$. Since the mesh of P is less than δ we know that $M_i - m_i < \frac{1}{j}$ if $[x_{i-1}, x_i] \cap K \neq \emptyset$, so $\sum_{\{i\in\mathbb{N}|[x_{i-1},x_i]\cap K \neq \emptyset\}} (M_i - m_i)(x_i - x_{i-1}) < \frac{b-a}{j} < \frac{\epsilon}{2}$.

By Theorem 7.72, we know that $\sum_{\{i\in\mathbb{N}|[x_{i-1},x_i]\cap K = \emptyset\}} (x_i - x_{i-1}) < \frac{\epsilon}{4M}$ since F covers the union of all intervals $[x_{i-1}, x_i]$ which do not intersect K, so $\sum_{\{i\in\mathbb{N}|[x_{i-1},x_i]\cap K = \emptyset\}} (M_i - m_i)(x_i - x_{i-1}) < \frac{\epsilon}{4M}(2M) = \frac{\epsilon}{2}$. Hence, $U(f,P) - L(f,P) < \epsilon$ and f is integrable.

□

Note: The name "Lebesgue Characterization of Riemann Integrability" is descriptive, and a name we may refer to, but it does not appear to be an official name for the preceding theorem that is normally used in the literature. It is referred to as "Lebesgue Criterion of Riemann Integrability" or the "Riemann-Lebesgue Theorem," but there does not appear to be a consistently used name for the theorem that is universally preferred.

Wallis's Formula:

Wallis's Formula is very much optional, but it makes some common definite integrals quick to evaluate and will simply our work in some later examples and exercises. It uses a reduction formula, so we will derive some of those first, using integration by parts.

Theorem 7.74. *Trigonometric integral reduction formulas. Let n, m be positive integer powers greater than or equal to two. Then:*

(1) $\int \sec^n(x)dx = \frac{1}{n-1}\sec^{n-2}(x)\tan(x) + \frac{n-2}{n-1}\int \sec^{n-2}(x)dx$

(2) $\displaystyle\int \csc^n(x)dx = -\frac{1}{n-1}\csc^{n-2}(x)\cot(x) + \frac{n-2}{n-1}\int \csc^{n-2}(x)dx$

(3) $\displaystyle\int \cos^n(x)dx = \frac{1}{n}\cos^{n-1}(x)\sin(x) + \frac{n-1}{n}\int \cos^{n-2}(x)dx$

(4) $\displaystyle\int \sin^n(x)dx = -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \frac{n-1}{n}\int \sin^{n-2}(x)dx$

(5) $\displaystyle\int \sin^n(x)\cos^m(x)dx = \frac{1}{n+1}\cos^{m-1}(x)\sin^{n+1}(x)+\frac{m-1}{n+1}\int \sin^{n+2}(x)\cos^{m-2}(x)dx$

(6) $\displaystyle\int \sin^n(x)\cos^m(x)dx = -\frac{1}{m+1}\sin^{n-1}(x)\cos^{m+1}(x)+\frac{n-1}{m+1}\int \sin^{n-2}(x)\cos^{m+2}(x)dx$

(7) $\displaystyle\int \tan^n(x)dx = \frac{1}{n}\tan^{n-1}(x) - \int \tan^{n-2}(x)dx$

(8) $\displaystyle\int \cot^n(x)dx = -\frac{1}{n}\cot^{n-1}(x) - \int \cot^{n-2}(x)dx$

Proof. (1) We use integration by parts, with the part to be differentiated $\sec^{n-2}(x)$ and the factor to be integrated $\sec^2(x)$. This gives us:

$\displaystyle\int \sec^n(x)dx = \sec^{n-2}(x)\tan(x) - \int (n-2)\sec^{n-3}(x)\sec(x)\tan(x)\tan(x)dx$. Then, recalling that $\tan^2(x) = \sec^2(x) - 1$ in the last integral, this becomes:

$$\int \sec^n(x)dx = \sec^{n-2}(x)\tan(x) - \int (n-2)\sec^{n-2}(x)(\sec^2(x)-1)dx =$$

$$\int \sec^n(x)dx = \sec^{n-2}(x)\tan(x) + (n-2)\int \sec^{n-2}(x)dx - (n-2)\int \sec^n(x)dx$$

$$\int \sec^n(x)dx + (n-2)\int \sec^n(x)dx = \sec^{n-2}(x)\tan(x) + (n-2)\int \sec^{n-2}(x)dx$$

so $\displaystyle (n-1)\int \sec^n(x)dx = \sec^{n-2}(x)\tan(x) + (n-2)\int \sec^{n-2}(x)dx$

and hence $\displaystyle\int \sec^n(x)dx = \frac{1}{n-1}\sec^{n-2}(x)\tan(x) + \frac{n-2}{n-1}\int \sec^{n-2}(x)dx.$

(2) Using parts again with the integrated factor $dv = \csc^2(x)$ and the differentiated factor $u = \csc^{n-2}(x)$ we obtain:

$\displaystyle\int \csc^n(x)dx = -\csc^{n-2}(x)\cot(x) - \int (n-2)\csc^{n-3}(x)(-\csc(x)\cot(x))(-\cot(x))dx.$

Then, recalling that $\cot^2(x) = \csc^2(x) - 1$ in the last integral, this becomes:

$\displaystyle\int \csc^n(x)dx = -\csc^{n-2}(x)\cot(x) - \int (n-2)\sec^{n-2}(x)(\csc^2(x)-1)dx$ so

$$\int \csc^n(x)dx = -\csc^{n-2}(x)\cot(x) + (n-2)\int \csc^{n-2}(x)dx - (n-2)\int \csc^n(x)dx$$

$$(n-1)\int \csc^n(x)dx = -\csc^{n-2}(x)\cot(x) + (n-2)\int \csc^{n-2}(x)dx$$

$$\int \csc^n(x)dx = -\frac{1}{n-1}\csc^{n-2}(x)\cot(x) + \frac{n-2}{n-1}\int \csc^{n-2}(x)dx$$

(3) Use parts again, setting $u = \cos^{n-1}(x)$ and $dv = \cos(x)$ which gives:

$$\int \cos^n(x)dx = \cos^{n-1}(x)\sin(x) - \int (n-1)\cos^{n-2}(x)(-\sin(x))(\sin(x))dx$$

Using the identity $\sin^2(x) = 1 - \cos^2(x)$ gives:

$\int \cos^n(x)dx = \cos^{n-1}(x)\sin(x) + \int (n-1)\cos^{n-2}(x)(1-\cos^2(x))dx,$
$\int \cos^n(x)dx = \cos^{n-1}(x)\sin(x) + (n-1)\int \cos^{n-2}(x)dx - (n-1)\int \cos^n(x))dx,$
$(n-1)\int \cos^n(x)dx + \int \cos^n(x)dx = \cos^{n-2}(x)\sin(x)dx + (n-1)\int \cos^{n-2}(x)dx,$
so $n\int \cos^n(x)dx = \cos^{n-1}(x)\sin(x) + (n-1)\int \cos^{n-2}(x)dx$. Thus,
$\int \cos^n(x)dx = \frac{1}{n}\cos^{n-1}(x)\sin(x) + \frac{n-1}{n}\int \cos^{n-2}(x)dx$

(4) Use parts, setting $u = \sin^{n-1}(x)$ and $dv = \sin(x)$ which gives:
$\int \sin^n(x)dx = \sin^{n-1}(x)(-\cos(x)) - \int (n-1)\sin^{n-2}(x)(\cos(x))(-\cos(x))dx.$
Using the identity $\cos^2(x) = 1 - \sin^2(x)$ on the last integral gives:
$\int \sin^n(x)dx = -\sin^{n-1}(x)\cos(x) + \int (n-1)\sin^{n-2}(x)(1-\sin^2(x))dx$, so
$\int \sin^n(x)dx = -\sin^{n-1}(x)\cos(x) + (n-1)\int \sin^{n-2}(x)dx - (n-1)\int \sin^n(x)dx,$
$n\int \sin^n(x)dx = -\sin^{n-1}(x)\cos(x) + (n-1)\int \sin^{n-2}(x)dx$. Thus,
$\int \sin^n(x)dx = -\frac{1}{n}\sin^{n-1}(x)\cos(x) + \frac{n-1}{n}\int \sin^{n-2}(x)dx$

(5) Use parts directly with the factor to be integrated being $\sin^n(x)\cos(x)$ and the factor to be differentiated being $\cos^{m-1}(x)$. A single use of parts gives us $\int \sin^n(x)\cos^m(x)dx = \frac{1}{n+1}\cos^{m-1}(x)\sin^{n+1}(x) + \frac{m-1}{n+1}\int \sin^{n+2}(x)\cos^{m-2}(x)dx$ as desired.

(6) Use parts again, with the factor to be integrated being $\cos^m(x)\sin(x)$ and the factor to be differentiated being $\sin^{n-1}(x)$. Using parts gives us $\int \sin^n(x)\cos^m(x)dx = -\frac{1}{m+1}\sin^{n-1}(x)\cos^{m+1}(x) + \frac{n-1}{m+1}\int \sin^{n-2}(x)\cos^{m+2}(x)dx.$

(7) The last two formulas don't require parts to derive. We just use the formulas $\tan^2(x) = \sec^2(x) - 1$ and $\cot^2(x) = \csc^2(x) - 1)$ as follows:
$\int \tan^n(x)dx = \int \tan^{n-2}(x)\tan^2(x)dx = \int \tan^{n-2}(x)\sec^2(x)dx - \int \tan^{n-2}(x)dx.$
Setting $u = \tan(x)$, $du = \sec^2(x)$ for the first integral, giving:
$\int \tan^n(x)dx = \int u^{n-2}du - \int \tan^{n-2}(x)dx$, so
$\int \tan^n(x)dx = \frac{\tan^{n-1}(x)}{n-1} - \int \tan^{n-2}(x)dx$

(8) The last formula is derived similarly.
$\int \cot^n(x)dx = \int \cot^{n-2}(x)\cot^2(x)dx = \int \cot^{n-2}(x)\csc^2(x)dc - \int \cot^{n-2}(x)dx.$
Setting $u = \cot(x)$, $du = -\csc^2(x)$ for the first integral, giving:
$\int \cot^n(x)dx = -\int u^{n-2}du - \int \cot^{n-2}(x)dx$, so

$$\int \cot^n(x)dx = -\frac{\cot^{n-1}(x)}{n-1} - \int \cot^{n-2}(x)dx$$

□

Theorem 7.75. *Wallis's Formula. Let m, n be non-negative integers. Then the integral $\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx = \frac{(n-1)(n-3)(n-5)...(1)(m-1)(m-3)(m-5)...(1)}{(n+m)(n+m-2)(n+m-4)...(1)}$ times $\frac{\pi}{2}$ if both n and m are even. Furthermore, for any integers $k < j$, if the number of intervals of the form $(i\frac{\pi}{2}, (i+1)\frac{\pi}{2})$ for $k \leq i < j$ for which $\sin^n(x)\cos^m(x)$ is positive on $(i\frac{\pi}{2}, (i+1)\frac{\pi}{2})$ is P and the number of intervals of the form $(i\frac{\pi}{2}, (i+1)\frac{\pi}{2})$ for $k \leq i < j$ for which $\sin^n(x)\cos^m(x)$ is negative on $(i\frac{\pi}{2}, (i+1)\frac{\pi}{2})$ is N then $\int_{\frac{k\pi}{2}}^{\frac{j\pi}{2}} \sin^n(x)\cos^m(x)dx = (P-N)\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx.$*

Proof. We use Theorem 7.74 part (5) on the integral $\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx$ we get $\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx = \cos^{m-1}(x)\frac{\sin^{n+1}(x)}{n+1}\Big|_0^{\frac{\pi}{2}} + \frac{m-1}{n+1}\int_0^{\frac{\pi}{2}} \sin^{n+2}(x)\cos^{m-2}(x)dx$ which is just equal to the integral $\frac{m-1}{n+1}\int_0^{\frac{\pi}{2}} \sin^{n+2}(x)\cos^{m-2}(x)dx$, assuming that $m-1$ is positive. Using (5) again on this integral with $n+2$ and $m-2$ as the new powers gives us that the integral is equal to $\frac{m-3}{n+3}\frac{m-1}{n+1}\int_0^{\frac{\pi}{2}} \sin^{n+2}(x)\cos^{m-2}(x)dx$, assuming that $m-3$ is still positive. We iterate this process until the power of cosine is either zero or one. In the case where m is even we are able to repeat this process $\frac{m}{2}$ times, leaving us with $\frac{(m-1)(m-3)...(1)}{(n+m-1)(n+m-3)...(n+3)(n+1)}\int_0^{\frac{\pi}{2}} \sin^{n+m}(x)dx$. If m is odd then we are only able to perform this reduction $\frac{m-1}{2}$ times, leaving us with $\frac{(m-1)(m-3)...(1)}{(n+m-3)...(n+3)(n+1)}\int_0^{\frac{\pi}{2}} \sin^{n+m-2}(x)\cos(x)dx$. Then, setting $u = \sin(x)$ and $du = \cos(x)$ the last integral becomes $\int_0^1 u^{n+m-2}du = \frac{1}{n+m-1}$. Multiplying this by the preceding integral gives $\frac{(m-1)(m-3)...(1)}{(n+m-1)(n+m-3)...(n+3)(n+1)}$ as desired.

Assuming that m is even, we continue, using formula (4), to give $\int_0^{\frac{\pi}{2}} \sin^{n+m}(x)dx = -\frac{1}{n+m}\sin^{n+m-1}(x)\cos(x)\Big|_0^{\frac{\pi}{2}} + \frac{n+m-1}{n+m}\int \sin^{n+m-2}(x)dx$. Assuming that $n+m-1$ is positive this becomes $\frac{n+m-1}{n+m}\int \sin^{n+m-2}(x)dx$. In the case where $n+m$ is even we are able to repeat this process $\frac{n+m}{2}$ times, finally arriving at the integral $\frac{(n+m-1)(n+m-3)...(1)}{(n+m)(n+m-2)...(1)}\int_0^{\frac{\pi}{2}} 1dx$. Thus, the original integral if both n, m are even is $\frac{(m-1)(m-3)...(1)}{(n+m-1)(n+m-3)...(n+3)(n+1)}\frac{(n+m-1)(n+m-3)...(1)}{(n+m)(n+m-2)...(1)}\frac{\pi}{2}$. The denoiminator of the first fraction cancels with all terms down to the $(n-1)$ factor in the numerator of the first fraction, leaving us with Wallis's formula, $\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx =$ $\frac{(n-1)(n-3)(n-5)...(1)(m-1)(m-3)(m-5)...(1)}{(n+m)(n+m-2)(n+m-4)...(1)}\frac{\pi}{2}$, as desired.

In the event that m is even but n is odd, we can only repeat this process $\frac{n+m-1}{2}$ times, leaving us with $\int_0^{\frac{\pi}{2}} \sin^{n+m}(x) = \frac{(n+m-1)(n+m-3)...(1)}{(n+m)(n+m-2)...(1)}\int_0^{\frac{\pi}{2}} \sin(x)dx$. Since $\int_0^{\frac{pi}{2}} \sin(x)dx = 1$, we get that $\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx =$ $\frac{(m-1)(m-3)...(1)}{(n+m-1)(n+m-3)...(n+3)(n+1)}\frac{(n+m-1)(n+m-3)...(1)}{(n+m)(n+m-2)...(1)}(1) =$ $\frac{(n-1)(n-3)(n-5)...(1)(m-1)(m-3)(m-5)...(1)}{(n+m)(n+m-2)(n+m-4)...(1)}$, as desired.

The last part of the theorem follows from a substitution and a trigonometric identity. We focus on integrating over a particular integer multiple of $\frac{\pi}{2}$ to its immediate successor times $\frac{\pi}{2}$ first, integrating $\int_{\frac{i\pi}{2}}^{\frac{(i+1)\pi}{2}} \sin^n(x)\cos^m(x)dx$. We first make the substitution $u = x - \frac{i\pi}{2}$. Then $du = dx$ and $x = u + \frac{i\pi}{2}$. The integral then becomes $\int_0^{\frac{\pi}{2}} \sin^n(u+\frac{i\pi}{2})\cos^m(u+\frac{i\pi}{2})du$. Using the sine and cosine sum of angles formulas, we see that $\sin(u+\frac{i\pi}{2}) = \sin(u)\cos(\frac{i\pi}{2}) + \cos(u)\sin(\frac{i\pi}{2})$, and $\cos(u+\frac{i\pi}{2}) = \cos(u)\cos(\frac{i\pi}{2}) - \sin(u)\sin(\frac{i\pi}{2})$. If i is odd then these simplify to $\sin(u+\frac{i\pi}{2}) = \pm\cos(u)$ and $\cos(u+\frac{i\pi}{2}) = \pm\sin(u)$. If i is even then they simplify to $\sin(u+\frac{i\pi}{2}) = \pm\sin(u)$ and $\cos(u+\frac{i\pi}{2}) = \pm\cos(u)$. Hence, in all possible cases the integrand is $\pm\sin^n(u)\cos^m(u)$ or $\pm\sin^m(u)\cos^n(u)$. By Wallis's formula (over $[0,\frac{\pi}{2}]$) this tells us that $\int_{\frac{i\pi}{2}}^{\frac{(i+1)\pi}{2}} \sin^n(x)\cos^m(x)dx = \pm\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx$, where the integral will be positive if $\sin^n(x)\cos^m(x) > 0$ on $(i\frac{\pi}{2},(i+1)\frac{\pi}{2})$ and negative if $\sin^n(x)\cos^m(x) < 0$ on $(i\frac{\pi}{2},(i+1)\frac{\pi}{2})$.

If we integrate $\int_{\frac{k\pi}{2}}^{\frac{j\pi}{2}} \sin^n(x)\cos^m(x)dx$ for integers $k < j$ then we can separate the integral into the sum of the integrals over the P intervals $(i\frac{\pi}{2},(i+1)\frac{\pi}{2})$ where $\sin^n(x)\cos^m(x) > 0$ and $k \leq i < j$, which equals $P\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx$, minus the sum of the integrals over the N intervals $(i\frac{\pi}{2},(i+1)\frac{\pi}{2})$ where $\sin^n(x)\cos^m(x) > 0$ and $k \leq i < j$, which equals $N\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx$, so $\int_{\frac{k\pi}{2}}^{\frac{j\pi}{2}} \sin^n(x)\cos^m(x)dx = (P-N)\int_0^{\frac{\pi}{2}} \sin^n(x)\cos^m(x)dx$.

□

7.9 Exercises for Supplementary Materials for One Variable

Exercise 7.1. *Let $m, n \in \mathbb{N}$. Then either $m^{\frac{1}{n}} \in \mathbb{N}$ or $m^{\frac{1}{n}}$ is irrational.*

Exercise 7.2. *Euclidean Algorithm (main premise). Let a, b be positive integers so that $a \leq b$ and q is the largest integer so that $b - aq \geq 0$. Then the greatest common divisor of a and b is the same as the greatest common divisor of a and $b - aq$.*

Exercise 7.3. *Let m, n be natural numbers then there are integers s, t so that $\dfrac{m}{n} = \dfrac{s}{t}$ so that the prime factor decompositions of s and t share any prime factors (such a fraction $\dfrac{s}{t}$ is said to be written in reduced terms). If p is a prime factor of m and is not a prime factor of n and $\dfrac{m}{n}$ is an integer then p is a prime factor of the integer $\dfrac{m}{n}$.*

Exercise 7.4. *Fermat's Little Theorem. Let p be prime. Then $a^p - a$ is divisible by p for every natural number a.*

Exercise 7.5. *If A and B are countable sets then $A \times B$ is countable.*

Exercise 7.6. *Let $a < b$. Then there is an irrational number r so that $a < r < b$.*

Exercise 7.7. *Let $f : E \to \mathbb{R}$. Then f is continuous if and only if for every closed set $A \subset \mathbb{R}$, the set $f^{-1}(A)$ is closed in E.*

Exercise 7.8. *Let S be uncountable. Then S has a limit point.*

Exercise 7.9. *Let $f : [a, b] \to \mathbb{R}$ be integrable. Let $g : [a, b] \to \mathbb{R}$ and let $F = \{x_1, x_2, ..., x_n\}$ be a finite subset of $[a, b]$, where $g(x) = f(x)$ for all $x \in [a, b] \setminus F$. Then $\inf_a^b g = \int_a^b f$.*

Hints:

Hint to Exercise 7.1. *Let $m, n \in \mathbb{N}$. Then either $m^{\frac{1}{n}} \in \mathbb{N}$ or $m^{\frac{1}{n}}$ is irrational.*

Suppose by way of contradiction that $\dfrac{p}{q} = m^{\frac{1}{n}}$, where $p \in \mathbb{N}$ and $q \in \mathbb{N}$ and $\dfrac{p}{q}$ is in reduced terms with $q > 1$. Explain why it is impossible for p^n to be an integer multiple of q^n.

Hint to Exercise 7.2. *Euclidean Algorithm (main premise). Let a, b be positive integers so that $a \le b$ and q is the largest integer so that $b - aq \ge 0$. Then the greatest common divisor of a and b is the same as the greatest common divisor of a and $b - aq$.*

First, assume that D is a common divisor of a and $b - aq$. Show that D is also a divisor of b using the fact that $a = mD$ and $b - aq = nD$ for some integers n, m. Then show that any common divisor of a and b is also a divisor of a and $b - aq$.

Hint to Exercise 7.3. *Let m, n be natural numbers then there are integers s, t so that $\dfrac{m}{n} = \dfrac{s}{t}$ so that the prime factor decompositions of s and t share any prime factors (such a fraction $\dfrac{s}{t}$ is said to be written in reduced terms). If p is a prime factor of m and is not a prime factor of n and $\dfrac{m}{n}$ is an integer then p is a prime factor of the integer $\dfrac{m}{n}$.*

Use the Fundamental Theorem of Arithmetic.

Hint to Exercise 7.4. *Fermat's Little Theorem. Let p be prime. Then $a^p - a$ is divisible by p for every natural number a.*

Use induction and the definition of divisibility. Recall that what it means for $a^p - a$ to be divisible by p is that $a^p - a = mp$ for some natural number m. Use the Binomial Theorem and explain why, for a prime number p, it is true that the factor p in the numerator of $\dfrac{p!}{i!(p-i)!}$ does not cancel with any factor of the denominator, assuming $1 \le i \le p - 1$. You may wish to use the Fundamental Theorem of Arithmetic for that.

Hint to Exercise 7.5. *If A and B are countable sets then $A \times B$ is countable.*

Use the fact that the union of countably many countable sets is countable.

Hint to Exercise 7.6. *Let $a < b$. Then there is an irrational number r so that $a < r < b$.*

If all the points between a and b are rational, then explain why the set of such points is countable (which contradicts a theorem).

Hint to Exercise 7.7. *Let $f : E \to \mathbb{R}$. Then f is continuous if and only if for every closed set $A \subset \mathbb{R}$, the set $f^{-1}(A)$ is closed in E.*

Try using Theorem 7.45.

Hint to Exercise 7.8. *Let S be uncountable. Then S has a limit point.*

Use the fact that the countable union of countable sets is countable to show that a bounded interval contains uncountably many points of S.

Hint to Exercise 7.9. *Let $f : [a,b] \to \mathbb{R}$ be integrable. Let $g : [a,b] \to \mathbb{R}$ and let $F = \{x_1, x_2, ..., x_n\}$ be a finite subset of $[a,b]$, where $g(x) = f(x)$ for all $x \in [a,b] \setminus F$. Then $\displaystyle\int_a^b g = \int_a^b f$.*

Use the Lebesgue Characterization of Riemann Integrability to show that g is integrable and then use the characterization of integral in terms of a sequence of Riemann sums to show the integrals are equal.

Solutions:

Solution to Exercise 7.1. *Let $m, n \in \mathbb{N}$. Then either $m^{\frac{1}{n}} \in \mathbb{N}$ or $m^{\frac{1}{n}}$ is irrational.*

Proof. Suppose that $\frac{p}{q} = m^{\frac{1}{n}}$, where $p \in \mathbb{N}$ and $q \in \mathbb{N}$ and $\frac{p}{q}$ is in reduced terms with $q > 1$. Then $\frac{p^n}{q^n} = m$. Since p and q have no common prime factors, p^n and q^n have no common prime factors by the Fundamental Theorem of Arithmetic. This means that p^n cannot be a multiple of q^n and in particular p^n is not equal to $m(q^n)$. This is a contradiction. □

Solution to Exercise 7.2. *Euclidean Algorithm (main premise). Let a, b be positive integers so that $a \leq b$ and q is the largest integer so that $b - aq \geq 0$. Then the greatest common divisor of a and b is the same as the greatest common divisor of a and $b - aq$.*

Proof. Let $d \in \mathbb{N}$. First, assume that d divides a and b. Then for some integers m, n it is true that $a = md$ and $b = nd$, which means that $b - aq = d(n - mq)$ and therefore d divides $b - aq$.

Suppose that d divides both a and $b - aq$. Then for some integers n, m we know that $b - aq = dm$ and $a = dn$. Thus, $b = d(m + nq)$ which means that d divides both a and b. Since the divisors of a and b are the same as the divisors of $b - aq$ it follows that both have the same greatest common divisor. □

Solution to Exercise 7.3. *Let m, n be natural numbers then there are integers s, t so that $\frac{m}{n} = \frac{s}{t}$ so that the prime factor decompositions of s and t share any prime factors (such a fraction $\frac{s}{t}$ is said to be written in reduced terms). If p is a prime factor of m and is not a prime factor of n and $\frac{m}{n}$ is an integer then p is a prime factor of the integer $\frac{m}{n}$.*

Proof. Let $m = p_1^{m_1} p_2^{m_2} ... p_j^{m_j}$ and $n = q_1^{n_1} q_2^{n_2} ... q_k^{n_k}$ be the prime decompositions of m and n. Let $p_{i_1}, ..., p_{i_s}$ be the primes in the decomposition of m that do not appear in the decomposition for n, and let $q_{j_1}, q_{j_2}, ..., q_{j_t}$ be the prime numbers in the decomposition of n that do not appear in the decomposition for m. Let $p_{i_{s+1}}, p_{i_{s+2}}, ..., p_{i_j}$ be the prime numbers which appear in both decompositions. Then if $v_{i_{s+r}}$ is the power of $p_{i_{s+r}}$ in the numerator minus the power of the same prime in the decomposition of the denominator, we have $\frac{m}{n} = (p_{i_1} .. p_{i_s})(p_{i_{s+1}}^{v_{i_{s+1}}} p_{i_{s+2}}^{v_{i_{s+2}}} ..., p_{i_j}^{v_{i_j}})(\frac{1}{q_{j_1}^{n_{j_1}} q_{j_2}^{n_{j_2}} ... q_{j_t}^{n_{j_t}}})$. This can be written as a fraction where, depending on whether the net power $v_{i_{s+r}}$ of each shared prime term is positive or negative, we put $p_{i_{s+r}}^{v_{i_{s+r}}}$ in the factorization of the numerator (if the exponent is positive) or $-p_{i_{s+r}}^{v_{i_{s+r}}}$ in the factorization in the numerator (if the exponent is negative). There are then no prime numbers common to the factorizations of both numerator and denominator.

Note that if $\frac{m}{n}$ is an integer then if p is one of the prime numbers in the factorization of m which does not occur in the prime factorization of n, the reduced form described above could only be an integer if all primes q_i canceled with corresponding primes in the numerator, and the prime factorization of the resulting fraction in reduced terms would still have p (in the numerator), so it would follow that p is a factor of $\frac{m}{n}$.

□

Solution to Exercise 7.4. *Fermat's Little Theorem. Let p be prime. Then $a^p - a$ is divisible by p for every natural number a.*

Proof. We proceed by induction on a. First, note that $1^p - 1 = 0$ is divisible by every number, including p. Assume that $k^p - k = mp$ for some natural number m. Then $(k+1)^p - (k+1) = -k - 1 + \sum_{i=0}^{p} \binom{p}{i} k^i = k^p - k + \sum_{i=1}^{p-1} \binom{p}{i} k^i$. Note that if $1 \leq i \leq p-1$ then $\binom{p}{i} = \frac{p!}{(p-i)!i!}$, where p divides the numerator, but since no prime factor of the denominator is p or larger, p does not divide the denominator since p is prime. To see this, use the Fundamental Theorem of Arithmetic to factor the each of the integers in the list $i, i-1, i-2, ..., 1$ into their prime decompositions (where each prime in the decomposition of each number is less than p since a number in a decomposition cannot exceed the number it is in the decomposition for) and multiply them together to get a product of prime numbers equal to $(i-1)!$. Similarly, none of $p-i, p-i-1, ..., 1$ has p as a factor in its prime decomposition. Thus, the prime decomposition of $(p-i)!i!$ does not contain any prime factor equal to (or larger than) p. Hence, the factor of p in $p!$ does not cancel with any of the prime numbers in the denominator of the fraction $\frac{p!}{(p-i)!i!}$, which means that $\binom{p}{i}$ is divisible by p (as addressed in the preceding exercise). Thus, there are natural numbers $m_1, m_2, ..., m_{p-1}$ so that $(k+1)^p - (k+1) = mp + m_1 p + m_2 p + ... + m_{p-1} p$ where each $m_i p = \frac{p!}{(p-i)!i!} k^i$, which means that $(k+1)^p - (k+1)$ is divisible by p. The result then follows by induction.

□

Solution to Exercise 7.5. *If A and B are countable sets then $A \times B$ is countable.*

Proof. If A or B is empty then $A \times B$ is empty and therefore countable. Assume A and B are not empty. By Theorem 7.19, since A is countable there is an onto function $h : \mathbb{N} \to A$, which means that we can list $A = \{a_1, a_2, a_3, ...\}$. Similarly, we can write $B = \{b_1, b_2, b_3, ...\}$.

By definition of Cartesian product, we have $A \times B = \bigcup_{i=1}^{\infty} \{a_i\} \times \{b_1, b_2, b_3, ...\}$. For each $i \in \mathbb{N}$, there is an onto function $f : \mathbb{N} \to \{a_i\} \times \{b_1, b_2, b_3, ...\} \to B$ defined by $f(j) = (i, h(j))$ for each $j \in \mathbb{N}$. Hence, $\{a_i\} \times \{b_1, b_2, b_3, ...\}$ is countable for each $i \in \mathbb{N}$. Thus, $A \times B$ is a countable union of countable sets and is therefore countable.

□

Solution to Exercise 7.6. *Let $a < b$. Then there is an irrational number r so that $a < r < b$.*

Proof. Since we have shown that each set containing an open interval is uncountable, we know that (a, b) is uncountable. Since $\mathbb{Q} \cap (a, b) \subset \mathbb{Q}$ and $\mathbb{Q}$ is countable, we know that $\mathbb{Q} \cap (a, b)$ is countable since subset of a countable set is countable, so there are irrational numbers in (a, b).

□

Solution to Exercise 7.7. *Let $f : E \to \mathbb{R}$. Then f is continuous if and only if for every closed set $A \subset \mathbb{R}$, the set $f^{-1}(A)$ is closed in E.*

Proof. Assume f is continuous and let A be closed. By Theorem 7.45, we know that $f^{-1}(\mathbb{R} \setminus A)$ is open in E, which means that $E \setminus f^{-1}(\mathbb{R} \setminus A) = f^{-1}(A)$ is closed in E.

Assume for every closed set $A \subset \mathbb{R}$, the set $f^{-1}(A)$ is closed in E. Let U be open. Then $f^{-1}(U) = E \setminus f^{-1}(\mathbb{R} \setminus A)$, which is the complement of a closed set in E, which means that $f^{-1}(U)$ is open in E. Thus, by Theorem 7.45, we know that f is continuous. □

Solution to Exercise 7.8. *Let S be uncountable. Then S has a limit point.*

Proof. If each $E_i = \{[i, i+1] \cap S\}$, where $i \in \mathbb{Z}$, is countable, then by Theorem 7.23, it follows that $S = \bigcup_{i \in \mathbb{Z}} E_i$ is countable (which is impossible). Hence, for some integer m we know that $E_m = [m, m+1] \cap S$ is uncountable and therefore infinite, which means that E_m has a limit point p by Exercise 3.18, and therefore p is a limit point of S by Exercise 3.14. □

Solution to Exercise 7.9. *Let $f : [a, b] \to \mathbb{R}$ be integrable. Let $g : [a, b] \to \mathbb{R}$ and let F be a finite subset of $[a, b]$, where $g(x) = f(x)$ for all $x \in [a, b] \setminus F$. Then $\int_a^b g = \int_a^b f$.*

Proof. Let $D_f = \{x \in [a, b] | f \text{ is not continuous at } x\}$. $D_g = \{x \in [a, b] | g \text{ is not continuous at } x\}$. Let $p \in [a, b] \setminus F$ and let f be continuous at p. Since F is finite, it is a finite union of (trivial) closed intervals and therefore a finite union of closed sets, which is closed. Let $\epsilon > 0$. Choose $\delta_1 > 0$ so that if $|p - x| < \delta_1$ and $x \in [a, b]$ then $|f(p) - f(x)| < \epsilon$. Since F is closed, we can find δ_2 so that if $(p - \delta_2, p + \delta_2) \cap F = \emptyset$. If $\delta = \min(\delta_1, \delta_2)$ then if $|x - p| < \delta$ and $x \in [a, b]$ then $|g(x) - g(p)| = |f(x) - f(p)| < \epsilon$, so g is continuous at p. Hence $D_g \subseteq D_f \cup F$. Since f is integrable, we know that $\lambda(D_f) = 0$. Since F is finite, $\lambda(F) = 0$. Hence, by Theorem 7.62, we know that $\lambda(D_f \cup F) = 0$, so by Theorem 7.60, we know that $\lambda(D_g) = 0$, so g is integrable by Theorem 7.73.

Choose a sequence of partitions $\{P_n\}$ of $[a, b]$ so that $\{|P_n|\} \to 0$. Since F is finite, for each P_n we can choose a marking T_n so that $T_n \cap F = \emptyset$. Thus, $S_{T_n}(f, P_n) = S_{T_n}(g, P_n)$ for all $n \in \mathbb{N}$. Hence, by Theorem 6.7, we know that $\{S_{T_n}(f, P_n)\} \to \int_a^b f$ and therefore $\{S_{T_n}(g, P_n)\} \to \int_a^b f$, and hence $\to \int_a^b f = \int_a^b g$.

□

Chapter 8

Series

Definition 57

If $\{x_n\}$ is a sequence then the *nth partial sum* of this sequence is $s_n = \sum_{i=1}^{n} x_i$. The sequence of partial sums $\{s_n\}$ is the *series* $\sum_{n=1}^{\infty} x_n$. We also use $\sum_{n=1}^{\infty} x_n$ to refer to the point to which this series converges, depending on context.

Theorem 8.1. *Cauchy Convergence Criterion. The series* $\sum_{n=1}^{\infty} x_n$ *converges if and only if for every* $\epsilon > 0$ *there is an integer* k *so that if* $n > m \geq k$ *then* $|\sum_{i=m+1}^{n} x_i| < \epsilon$.

Proof. We know that the sequence of partial sums $\{s_n\} = \{\sum_{i=1}^{n} x_i\}$ converges if and only if it is a Cauchy sequence by Theorem 3.25, which is true if and only if for every $\epsilon > 0$ there is an integer k so that if $n > m \geq k$ then $|s_n - s_m| = |\sum_{i=m+1}^{n} x_i| < \epsilon$.

□

Theorem 8.2. *Let* k *be a non-negative integer and* $c \neq 0$ *and let* $\{x_n\}$ *be a sequence. Then* $\sum_{n=1}^{\infty} x_n$ *converges if and only if* $\sum_{n=k+1}^{\infty} cx_n$ *converges. Furthermore, if* $\sum_{n=1}^{\infty} x_n = L$ *then* $\sum_{n=k+1}^{\infty} cx_n = cL - \sum_{n=1}^{k} cx_n$.

Proof. Let $\{s_n\}$ be the sequence of partial sums of the sequence $\{x_n\}$ and note that $\{s_{n+k} - s_k\} = \sum_{n=k+1}^{\infty} x_n$. Then $\{s_n\} \to L$ if and only if $\{s_{n+k}\} \to L$ by Theorem 3.26, and $\{s_{n+k}\} \to$

L if and only if $\{s_{n+k} - s_k\} \to L - s_k$. Thus, $\{s_n\} \to L$ if any only if $\{c(s_{n+k} - s_k)\} \to c(L - s_k)$ by the product rule for sequence limits. In other words, $\sum_{n=k+1}^{\infty} cx_n = cL - \sum_{n=1}^{k} cx_n$.

□

Theorem 8.3. *Divergence Test. If $\sum_{n=1}^{\infty} x_n$ converges then $\{x_n\} \to 0$.*

Proof. Let $s_n = \sum_{i=1}^{n} x_i$ be a sequence converging to the point s_∞. Then $\{s_{n+1}\}$ is a subsequence of $\{s_n\}$ converging to s_∞ by Theorem 3.11. Hence the sequence $\{s_{n+1} - s_n\} = \{x_{n+1}\} \to s_\infty - s_\infty = 0$, so $\{x_n\} \to 0$.

□

Theorem 8.4. *Comparison Test. Let $\{a_n\}$ and $\{b_n\}$ be sequences of non-negative terms so that $a_n \leq b_n$ for all $n \in \mathbb{N}$. If $\sum_{n=1}^{\infty} b_n$ converges then $\sum_{n=1}^{\infty} a_n$ converges. If $\sum_{n=1}^{\infty} a_n$ diverges then $\sum_{n=1}^{\infty} b_n$ diverges.*

Proof. Let $A_n = \sum_{i=1}^{n} a_i$, and let $B_n = \sum_{i=1}^{n} b_i$. Since the terms of $\{a_n\}$ and $\{b_n\}$ are non-negative, it follows that $\{A_n\}$ and $\{B_n\}$ are non-decreasing sequences, and by The Monotone Convergence Theorem they converge if and only if they are bounded above. Thus, if $\sum_{n=1}^{\infty} b_n$ converges then $\{B_n\}$ is bounded above, so $\{A_n\}$ is bounded above since $A_n \leq B_n$ for each $n \in \mathbb{N}$, so $\sum_{n=1}^{\infty} a_n$ converges. Hence, if $\sum_{n=1}^{\infty} a_n$ diverges then $\sum_{n=1}^{\infty} b_n$ diverges.

□

The comparison test can be thought of as the strongest test for verifying the convergence of series sums of non-negative terms in the sense that every convergent series consisting of only positive terms is less than some other convergent series, so if you could find the right (larger) series to compare to and show this larger series is convergent then you could always show the smaller series is convergent. In practice, however, this is not always reasonable. Frequently, a good series to compare to is a p-series, addressed in the exercises.

Example 8.1. *Determine whether $\sum_{n=1}^{\infty} \frac{5 + \cos(n)}{n^3 + \ln(n+4)}$ converges (with justification).*

Solution. Since $5+\cos(n) < 6$ and $n^3+\ln(n+4) > n^3$ for all $n \in \mathbb{N}$ it follows that $\dfrac{5+\cos(n)}{n^3+\ln(n+4)} < \dfrac{6}{n^3}$ for all positive integers n. Since we know that the p-series $\displaystyle\sum_{n=1}^{\infty}\frac{1}{n^3}$ converges, it follows from Theorem 8.2 that $\displaystyle\sum_{n=1}^{\infty}\frac{6}{n^3}$ converges, so $\displaystyle\sum_{n=1}^{\infty}\frac{5+\cos(n)}{n^3+\ln(n+4)}$ converges by the Comparison Test.

□

Theorem 8.5. *Limit Comparison Test. Let $\{a_n\}$ and $\{b_n\}$ be sequences of non-negative terms so that $\displaystyle\lim_{n\to\infty}\frac{a_n}{b_n} = L > 0$. Then $\displaystyle\sum_{n=1}^{\infty}a_n$ converges if and only if $\displaystyle\sum_{n=1}^{\infty}b_n$ converges.*

Proof. Let $A_n = \displaystyle\sum_{i=1}^{n}a_i$, and let $B_n = \displaystyle\sum_{i=1}^{n}b_i$. We can choose a positive integer N so that if $n \geq N$ then $|\dfrac{a_n}{b_n} - L| < \dfrac{L}{2}$, and hence $\dfrac{L}{2} < \dfrac{a_n}{b_n} < \dfrac{3L}{2}$, which means that $\dfrac{b_nL}{2} < a_n < \dfrac{3Lb_n}{2}$ for all $n \geq N$. By Theorem 8.2, it follows that if $\displaystyle\sum_{n=N+1}^{\infty}b_n$ converges then $\displaystyle\sum_{n=N+1}^{\infty}\frac{3Lb_n}{2}$ converges, and so $\displaystyle\sum_{n=N+1}^{\infty}a_n$ converges by the comparison test. Similarly, if $\displaystyle\sum_{n=N+1}^{\infty}a_n$ converges then $\displaystyle\sum_{n=N+1}^{\infty}\frac{b_nL}{2}$ converges and so $\displaystyle\sum_{n=N+1}^{\infty}b_n$ converges. Hence, $\displaystyle\sum_{n=1}^{\infty}a_n$ converges if and only if $\displaystyle\sum_{n=1}^{\infty}b_n$ converges.

□

Theorem 8.6. *Geometric Series Convergence. Let $a_n = ar^{n-1}$ for each positive integer n. Then $s_n = \displaystyle\sum_{i=1}^{n}ar^{i-1} = \frac{a(1-r^n)}{1-r}$ and if $|r|<1$ then $\displaystyle\sum_{n=1}^{\infty}a_n = \frac{a}{1-r}$.*

Proof. Note that $rs_n = \displaystyle\sum_{i=1}^{n}ar$, so $s_n - rs_n = a - ar^n$, so $s_n = \displaystyle\sum_{i=1}^{n}ar^{i-1} = \frac{a(1-r^n)}{1-r}$. If $|r|<1$ then $\displaystyle\lim_{n\to\infty}r^n = 0$, so $\displaystyle\sum_{i=1}^{\infty}ar^{i-1} = \lim_{n\to\infty}\frac{a(1-r^n)}{1-r} = \frac{a}{1-r}$.

□

Definition 58

We define $\displaystyle\int_a^{\infty} f = \lim_{b\to\infty}\int_a^b f$ if this limit exists.

Theorem 8.7. *Let $F : [a, \infty)$ be a monotone function. Then $\lim_{x\to\infty} F(x) = L$ exists if and only if F is bounded. If F is non-decreasing then L is the least upper bound of the range of F. If F is decreasing then L is the greatest lower bound of the range of F. If F is not bounded and is non-decreasing then $\lim_{x\to\infty} F(x) = \infty$. If F is not bounded and is non-increasing then $\lim_{x\to\infty} F(x) = -\infty$.*

Proof. We first assume that F is non-decreasing. If F is not bounded then it is not bounded above (the range of F is bounded below by $F(a)$, let u be the least upper bound for F. Given any $\epsilon > 0$ it follows that there is some x_0 so that $F(x_0) > u - \epsilon$. Since F is increasing we know that if $x > x_0$ then $u - \epsilon < F(x_0) < F(x) \leq u$, so $|F(x) - u| < \epsilon$ which means that $\int_1^\infty f(x)dx = \lim_{x\to\infty} F(x) = u$.

If the range of F is not bounded above then given any number M there is some $x_0 \in [a, \infty)$ so that $F(x_0) > M$, so if $x \geq x_0$ then $F(x) > M$, which means that $\lim_{x\to\infty} F(x) = \infty$.

Finally, if F is non-increasing then $-F$ is non-decreasing. Since the range of F is bounded if and only if the range of $-F$ is bounded, it follows that $-F$ converges if and only if the range of F is bounded, in which case $\lim_{x\to\infty} -F(x) = -b$, the least upper bound of the range of $-F$, but then b is the greatest lower bound of the range of F and $\lim_{x\to\infty} F(x) = b$. Likewise, if the range of F is not bounded then it is not bounded below, and so the range of $-F$ is not bounded above, from which we conclude that $\lim_{x\to\infty} -F(x) = \infty$, so $\lim_{x\to\infty} F(x) = -\infty$.

□

Theorem 8.8. *The Integral Test and Integral Remainder Theorem. Let $f : [1, \infty) \to (0, \infty)$ be a non-increasing function, where $f(i) = a_i$ for each $i \in \mathbb{N}$. Then the series $\sum_{i=1}^{n} a_i$ converges if and only if $\int_1^\infty f(x)dx$ converges. Furthermore, if we let $s_n = \sum_{i=1}^{n} a_i$ for each $n \in \mathbb{N}$ then if $L = \sum_{i=1}^{\infty} a_i$ then for any natural number m it follows that $s_m + \int_{m+1}^\infty f(x)dx \leq L \leq s_m + \int_m^\infty f(x)dx$.*

Proof. First, note that $F(x) = \int_1^x f(t)dt$ is defined by Exercise 6.6 since f is monotone, and F is increasing because if $1 \leq a < b$ it follows that $f(x) \geq f(b) > 0$ on $[a, b]$ which means that $F(b) - F(a) = \int_a^b f(x)dx \geq f(b)(b - a) > 0$. Likewise, since each $a_i > 0$, if we set $s_n = \sum_{i=1}^{n} a_i$ then $\{s_n\}$ is increasing since each $a_i > 0$ so for $n > m$ we have $s_n - s_m = \sum_{i=m+1}^{n} a_i > 0$.

Next, we note that, for each natural number i it is true that $a_i \geq f(x) \geq a_{i+1}$ for all $x \in [i, i+1]$. From this it follows that if $j, k \in \mathbb{N}$ so that $j < k$ then $\sum_{i=j}^{k-1} a_i \geq \sum_{i=j}^{k-1} \int_i^{i+1} f(x)dx = \int_j^k f(x)dx = F(k) - F(j)$. Likewise, it follows that $\sum_{i=j+1}^{k} a_i \leq \sum_{i=j}^{k-1} \int_i^{i+1} f(x)dx = \int_j^k f(x)dx = F(k) - F(j)$.

We then observe that if $\{s_n\} \to L$ then by the Monotone Convergence Theorem, it follows that L is the least upper bound for $\{s_n\}$. Since $\int_1^x f(t)dt \leq s_n < L$ for any $n \in \mathbb{N}$ that exceeds x, it follows that $F(x) < L$ for all $x \in [0, \infty)$ which means that the range of F is bounded and has a least upper bound u. By Theorem 8.7, $\lim_{x\to\infty} F(x) = \int_1^\infty f(x)dx = u$.

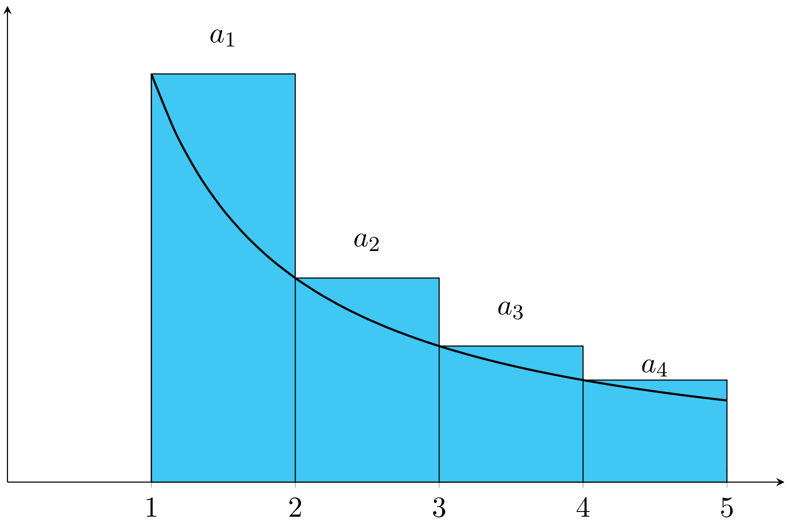
Conversely, if $\{s_n\}$ is divergent then $\{s_n\}$ is not bounded above, which means that given any number M there is some $k \in \mathbb{N}$ so that $s_k > a_1 + M$. It follows that if $x \geq k$ then $F(x) \geq F(k) \geq \sum_{i=2}^{k} a_i = s_k - a_1 > M$, so F is unbounded. Thus, by Theorem 8.7, $\int_1^\infty f(x)dx = \infty$.

Finally, let $m \in \mathbb{N}$ and let $\{s_n\} \to L$. Then $\sum_{i=m+1}^{\infty} a_i = L - s_m$. For each $n > m$ we also know that $\sum_{i=m+1}^{n-1} a_i \geq \int_{m+1}^{n} f(x)dx$ which implies that $\sum_{i=m+1}^{\infty} a_i \geq \int_{m+1}^{\infty} f(x)dx$, so $L = s_m + \sum_{i=m+1}^{\infty} a_i \geq s_m + \int_{m+1}^{\infty} f(x)dx$ by the Comparison Theorem. Likewise, since $\sum_{i=m+1}^{n} a_i \leq \int_m^n f(x)dx$ it follows from the Comparison Theorem that $L = s_m + \sum_{i=m+1}^{\infty} a_i \leq s_m + \int_m^\infty f(x)dx$.

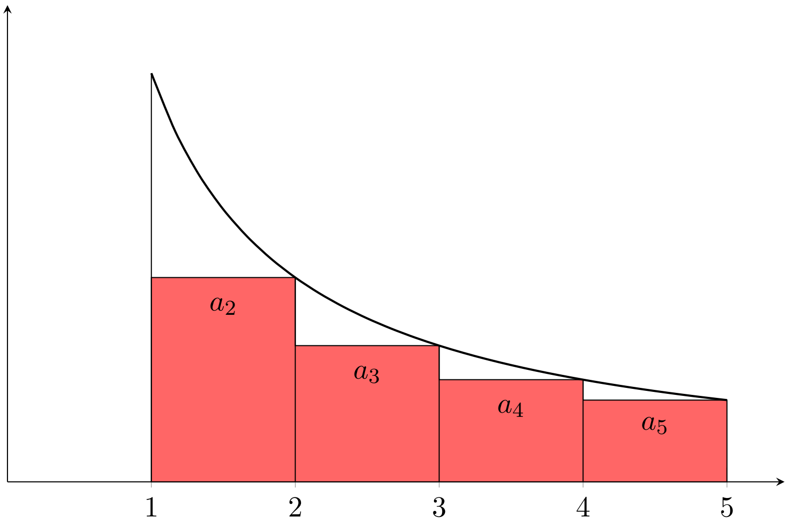
□

The following graph illustrates the argument for the proof of the Integral Test.

Area Under f Less than Sum



Area Under f More than Sum



Another way of viewing the remainder is as follows:

Theorem 8.9. *If $a_n = f(n)$, where f is a decreasing continuous function so that $\int_1^\infty f(x)dx$ converges, then for each natural number n, if $s_n = \sum_{i=1}^{n} a_i$ then $s_n + \dfrac{\int_n^\infty f(x)dx + \int_{n+1}^\infty f(x)dx}{2}$ is within $\dfrac{\int_n^\infty f(x)dx - \int_{n+1}^\infty f(x)dx}{2}$ of the series sum $s = \sum_{i=1}^{\infty} a_i$.*

Proof. By the preceding theorem, we know that $s \in [s_n + \int_{n+1}^{\infty} f(x)dx, s_n + \int_n^{\infty} f(x)dx]$, which means that the midpoint $s_n + \frac{\int_n^{\infty} f(x)dx + \int_n^{\infty} f(x)dx}{2}$ of this interval is within a distance of half its length from s, which is $\frac{\int_n^{\infty} f(x)dx - \int_{n+1}^{\infty} f(x)dx}{2}$. □

We should also point out that if we don't mind adding more terms then we have a second, simpler formula that $\sum_{i=1}^{\infty} a_i = s_n + R_n$, where $R_n = \sum_{i=n+1}^{\infty} a_i < \int_n^{\infty} f(x)dx$. If you estimate the series with a remainder that is bounded in this manner, the remainder is easier to find a bound for, but you must add more terms to achieve an estimate within a given error typically.

Also, observe that formula in the theorem would only yield a "$\leq$" sign, but if f is strictly decreasing we note that over a given interval $[i, \frac{i}{2}]$ it is the case that $f(x) \geq f(i+\frac{1}{2}) > f(i+1)$ which means that in fact $\int_i^{i+1} f(x)dx - f(i+1)$ is at least $\frac{1}{2}(f(i+\frac{1}{2}) - f(i+1)) > 0$, so the partial inequality signs could be replaced by strict inequality signs if f is strictly decreasing.

Definition 59

We say a series $\sum_{n=1}^{\infty} a_n$ converges *absolutely* if $\sum_{n=1}^{\infty} |a_n|$ converges. We say that $\sum_{n=1}^{\infty} a_n$ converges *conditionally* if it converges but does not converge absolutely.

Theorem 8.10. *If a series $\sum_{n=1}^{\infty} a_n$ converges absolutely then $\sum_{n=1}^{\infty} a_n$ converges.*

Proof. Let $\epsilon > 0$. Using the Cauchy Convergence Criterion, if $\sum_{n=1}^{\infty} a_n$ converges absolutely then we can find an integer k so that if $n > m \geq k$ then $|\sum_{i=m+1}^{n} |a_i|| = \sum_{i=m+1}^{n} |a_i| < \epsilon$, so by the triangle inequality $|\sum_{i=m+1}^{n} a_i| < \epsilon$, and therefore $\sum_{n=1}^{\infty} a_n$ converges.

□

Theorem 8.11. *Ratio Test. Let $\{a_n\}$ be a sequence so that $\lim_{n\to\infty}|\frac{a_{n+1}}{a_n}| = L$. Then $\sum_{n=1}^{\infty} a_n$ converges absolutely if $L < 1$, and $\sum_{n=1}^{\infty} a_n$ diverges if $L > 1$.*

Proof. If $L > 1$ then for sufficiently large n it follows that $|\frac{|a_{n+1}|}{|a_n|} - L| < L - 1$, so $\frac{|a_{n+1}|}{|a_n|} > 1$, from which we see that the terms $|a_n|$ are increasing, so $\sum_{n=1}^{\infty} a_n$ diverges by the Divergence Test.

Let $L < 1$. Choose $u \in (L, 1)$. Choose k so that if $n \geq k$ then $\frac{a_{n+1}}{a_n} < u$. Then $|\frac{a_{k+1}}{a_n}| < u$, so $|a_{k+1}| < u|a_k|$. Similarly, $|a_{k+2}| < u|a_{k+1}| < u^2 a_k$, and inductively we find that if $m = k + j$ then $|a_m| < u^j|a_k|$. Hence, $\sum_{n=k}^{\infty} a_n$ converges by comparison with the geometric series $\sum_{n=1}^{\infty} a_k u^{n-1}$, so $\sum_{n=1}^{\infty} a_n$ converges absolutely.

□

An quick consequence of the Ratio Test is the following theorem, which is sometimes helpful as well.

Theorem 8.12. *Let $\{a_n\}$ be a sequence so that $\lim_{n\to\infty}|\frac{a_{n+1}}{a_n}| = L$. Then $\{a_n\} \to 0$ if $L < 1$ and $\{|a_n|\} \to \infty$ if $L > 1$*

Proof. By the Ratio Test, if $L < 1$ then $\sum_{n=1}^{\infty} a_n$ converges so $\{a_n\} \to 0$ by the Divergence Test. If $L > 1$ then choose $1 < u < L$. For some $k \in \mathbb{N}$, if $n \geq k$ then $|\frac{a_{n+1}}{a_n}| > u$ which means that $|a_{k+m}| \geq u^m|a_k|$ so $\{|a_n|\} \to \infty$. □

Example 8.2. *Determine whether $\sum_{n=1}^{\infty} \frac{(-1)^n n^2}{n!}$ converges, with justification.*

Solution. We take the ratio of the $n + 1$st term over the nth term in absolute value and take the limit as the ratio approaches infinity. This is $\lim_{n\to\infty}\left|\frac{\frac{(-1)^(n+1)(n+1)^2}{(n+1)!}}{\frac{(-1)^n n^2}{n!}}\right|$. In the absolute value the $(-1)^n$ doesn't change anything to we ignore that. $\frac{(n+1)!}{n!} = n + 1$ and $\lim_{n\to\infty}\frac{(n+1)^2}{n^2} = 1$ since the numerator and denominator are both second degree with leading coefficient one. Hence, we get $\lim_{n\to\infty}\frac{1}{n+1} = 0 < 1$, so by the Ratio Test this series converges (absolutely).

□

When using the ratio test it is useful to notice that for any polynomial $P(x)$ it is true that $\lim\limits_{n\to\infty} \dfrac{P(n+1)}{P(n)} = 1$. This means, in particular, that the ratio test is useless for determining the convergence of series of rational functions of n since the limit will always be one, and factors of expressions which are rational functions will always correspond to a ratio converging to one, just as $\dfrac{(n+1)^2}{n^2}$ converged to one as n approached infinity in the example above. This means that polynomial factors of numerator or denominator will just multiply the resulting limit by one when the limit of the ratio of a_{n+1} and a_n is taken. We will prove that here, along with a second component that will help with the Root Test.

Theorem 8.13. *Let* $P(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$ *be a polynomial. Then* $\lim\limits_{n\to\infty} \dfrac{P(n+1)}{P(n)} = 1$ *and* $\lim\limits_{n\to\infty} |P(n)|^{\frac{1}{n}} = 1$.

Proof. Using the Binomial Theorem, $P(n+1) = a_n \sum\limits_{i=0}^{n} \binom{n}{i} x^i + a_{n-1} \sum\limits_{i=0}^{n-1} \binom{n}{i} x^i + ... + a_1 \sum\limits_{i=0}^{1} \binom{n}{i} x^i + a_0$. Hence, the leading coefficient is a_n and the degree of $P(n+1)$ is n in the variable n. Since $P(n)$ is also a polynomial of degree n with leading coefficient a_n, it follows that $\lim\limits_{n\to\infty} \dfrac{P(n+1)}{P(n)} = \dfrac{a_n}{a_n} = 1$.

Taking a logarithm we have $\ln(|P(n)|^{\frac{1}{n}}) = \dfrac{\ln(P(n)}{n}$. Since both numerator and denominator both approach infinity as n approaches infinity, we get $\lim\limits_{x\to\infty} \dfrac{\ln(P(x)}{x} = \lim\limits_{x\to\infty} \dfrac{P'(x)}{P(x)} = 0$ since $P'(x)$ is degree $n-1$ and $P(x)$ is degree n. Hence, $\lim\limits_{n\to\infty} |P(n)|^{\frac{1}{n}} = e^0 = 1$ by Theorems 7.33 and 7.40. □

Definition 60

Let $\{x_n\}$ be a sequence. Then $\limsup\{x_n\} = \lim\limits_{n\to\infty} \sup\{x_n, x_{n+1}, x_{n+2}, ...\}$. We also use the notation $\limsup x_n = \limsup\{x_n\}$ for brevity. Likewise, we define $\liminf\{x_n\} = \lim\limits_{n\to\infty} \inf\{x_n, x_{n+1}, x_{n+2}, ...\}$.

Note that $\{\sup\{x_n, x_{n+1}, x_{n+2}, ...\}\}$ is non-increasing and $\{\inf\{x_n, x_{n+1}, x_{n+2}, ...\}\}$ is non-decreasing, and thus $\limsup\{x_n\}, \liminf\{x_n\}$ always exist and are real numbers if $\{x_n\}$ is a bounded sequence (the $\limsup$ is always a real number if $\{x_n\}$ is bounded above and $\liminf$ is always a real number if $\{x_n\}$ is bounded below). If $\{x_n\}$ is not bounded above then $\sup\{x_n, x_{n+1}, x_{n+2}, ...\} = \infty$ for each $n \in \mathbb{N}$ in which case we say that $\limsup\{x_n\} = \infty$. Likewise, if $\{x_n\}$ is not bounded below then we say $\liminf\{x_n\} = -\infty$.

Theorem 8.14. *Root Test. Let $\{a_n\}$ be a sequence so that* $\limsup\{|a_n|^{\frac{1}{n}}\} = L$. *Then* $\sum_{n=1}^{\infty} a_n$ *converges absolutely if $L < 1$, and* $\sum_{n=1}^{\infty} a_n$ *diverges if $L > 1$. If* $\limsup\{|a_n|^{\frac{1}{n}}\} = \infty$ *then* $\sum_{n=1}^{\infty} a_n$ *also diverges.*

Proof. If $L > 1$ or $\{x_n\}$ is not bounded above then for sufficiently large n it follows that $|a_n|^{\frac{1}{n}} > 1$, so $|a_n| > 1$, so $\sum_{n=1}^{\infty} a_n$ diverges by the Divergence Test.

Let $L < 1$. Choose $u \in (L, 1)$. Choose k so that if $n \geq k$ then $|a_n|^{\frac{1}{n}} < u$. Then $|a_n| < u^n$ for $n \geq k$, and thus $\sum_{n=k}^{\infty} |a_n|$ converges by comparison with a geometric series so $\sum_{n=1}^{\infty} a_n$ converges absolutely.

□

Definition 61

For a sequence $\{a_n\}$ we will use the notation $A_{(n,m)} = \sum_{i=m}^{n} a_i$ for $n \geq m$. Essentially, the capital letter with index subscripts from the letter used to designate the sequence can be used to indicate sums from one index to another.

Theorem 8.15. *Abel's Formula. Let $\{a_i\}$ and $\{b_i\}$ be sequences and $n, m \in \mathbb{N}$ so that $n > m$. Then* $\sum_{i=m}^{n} a_i b_i = A_{(n,m)} b_n - \sum_{i=m}^{n-1} A_{i,m}(b_{i+1} - b_i)$.

Proof. Writing out the right side of the equation we have $A_{(n,m)} b_n - \sum_{i=m}^{n-1} A_{i,m}(b_{i+1} - b_i) =$ $b_n(a_m + a_{m+1} + ... + a_{n-1} + a_n) - b_n(a_m + a_{m+1} + ... + a_{n-1}) + b_{n-1}(a_m + a_{m+1} + ... + a_{n-1}) - b_{n-1}(a_m + a_{m+1} + ... + a_{n-2}) + ... + b_{m+1}(a_m + a_{m+1}) - b_{m+1}(a_{m+1}) + b_m a_m$. The only terms remaining after cancellation are $\sum_{i=m}^{n} a_i b_i$.

□

Theorem 8.16. *Dirichlet's Test. Let $\{b_k\}$ be a decreasing sequence whose terms approach zero, and let $\{a_n\}$ be a sequence whose partial sums are bounded. Then* $\sum_{i=1}^{\infty} a_i b_i$ *converges.*

Proof. We can choose $M > 2|\sum_{i=1}^{n} a_i| = 2|A_{n,1}|$ for all $n \in \mathbb{N}$ since $\{a_n\}$ is a sequence whose partial sums are bounded. Let $\epsilon > 0$ and choose $k \in \mathbb{N}$ so that if $m \geq k$ then $b_k < \frac{\epsilon}{M}$.

Then $|A_{(n,m)}| \leq |A_{m,1}| + |A_{n,m+1}| < M$ for all $n, m \in \mathbb{N}$ so that $n > m$. By Abel's formula, $|\sum_{i=m}^{n} a_i b_i| = |A_{(n,m)} b_n + \sum_{i=m}^{n-1} A_{i,m}(b_i - b_{i+1})| < M b_m < \epsilon$ if $m \geq k$ (since $b_i - b_{i+1}$ is always positive). Thus, $\sum_{i=1}^{\infty} a_i b_i$ converges by the Cauchy Convergence Criterion.

□

Definition 62

An *alternating series* is a series of the form $a_1 - a_2 + a_3 - a_4 + ...$ or the form $-a_1 + a_2 - a_3 + a_4 + ...$, where $a_1, a_2, a_3, ...$ are all positive numbers.

Theorem 8.17. *Alternating Series Test. Let $\{b_i\}$ be a decreasing sequence converging to zero. Then $\sum_{i=1}^{\infty}(-1)^i b_i$ converges.*

Proof. Setting $a_i = (-1)^i$ in Dirichlet's Test, we note that the partial sums of $\{a_i\}$ are bounded and therefore $\sum_{i=1}^{\infty} a_i b_i = \sum_{i=1}^{\infty}(-1)^i b_i$ converges.

□

While Dirichlet's Test has merit in and of itself, readers who are not interested in Abel's Theorem or Dirichlet's Test can simply prove the Alternating Series Test directly as follows. This proof also includes the remainder theorem.

Theorem 8.18. *Alternating Series Remainder Theorem. Let $\{a_i\}$ be a decreasing sequence converging to zero, and let $\sum_{i=1}^{\infty}(-1)^{i+1} a_i = L$. Then, for every $n \in \mathbb{N}$, it follows that L is between s_n and s_{n+1}.*

Proof. Since $a_1 > a_2 > a_3 > ...$, each $(a_{2n-1} - a_{2n}) > 0$ and each $(-a_{2n} + a_{2n+1}) < 0$, which means that $a_1 > a_1 + (-a_2 + a_3) > a_1 + (-a_2 + a_3) + (-a_4 + a_5)...$ and $(a_1 - a_2) < (a_1 - a_2) + (a_3 - a_4) < (a_1 - a_2) + (a_3 - a_4) + (a_5 - a_6)....$ Since $s_1 > s_3 > s_5 > ...$ and $s_2 < s_4 < s_6 < ...$, the odd indexed partial sums are a decreasing sequence and the even partial sums are an increasing sequence, both of which converge to L since they are subsequences of $\{s_n\}$ which we know converges to L. Hence, if n is even then $s_n < L < s_{n+1}$ and if n is odd then $s_n > L > s_{n+1}$.

□

Theorem 8.19. *Log Test. Let $\{a_n\}$ be a sequence of positive terms so that $\lim_{n\to\infty} \frac{\ln(a_n^{-1})}{\ln(n)} = L$. Then if $L > 1$, $\sum_{n=1}^{\infty} a_n$ converges and if $L < 1$ then $\sum_{n=1}^{\infty} a_n$ diverges. Furthermore, if $\liminf \frac{\ln(a_n^{-1})}{\ln(n)} = L > 1$ then $\sum_{n=1}^{\infty} a_n$ converges, and if $\limsup \frac{\ln(a_n^{-1})}{\ln(n)} = L < 1$ then $\sum_{n=1}^{\infty} a_n$ diverges*

Proof. First assume that $L > 1$. Choose $u \in (1, L)$. If either $\lim_{n\to\infty} \frac{\ln(a_n^{-1})}{\ln(n)} = L$ or $\liminf \frac{\ln(a_n^{-1})}{\ln(n)} = L$, then there is some $k \in \mathbb{N}$ so that if $n \geq k$ then $\frac{\ln(a_n^{-1})}{\ln(n)} > u$, which means that $\log_n a_n^{-1} > u$, so $a_n^{-1} > n^u$, and thus $a_n < \frac{1}{n^u}$. By the Comparison Test and Exercise 8.1 it follows that $\sum_{n=1}^{\infty} a_n$ converges.

Next, assume that $L < 1$ and choose $u \in (1, L)$. If either $\lim_{n\to\infty} \frac{\ln(a_n^{-1})}{\ln(n)} = L$ or $\limsup \frac{\ln(a_n^{-1})}{\ln(n)} = L$, then we can choose $k \in \mathbb{N}$ so that if $n \geq k$ then $\frac{\ln(a_n^{-1})}{\ln(n)} < u$. Then $a_n^{-1} < n^u$, so $a_n > \frac{1}{n^u}$. Thus, $\sum_{n=1}^{\infty} a_n$ diverges by the Comparison Test. □

Example 8.3. *Determine whether $\sum_{n=1}^{\infty} (\frac{1}{\ln(n)})^{\ln(n)}$ converges.*

Solution. Since $\lim_{n\to\infty} \frac{\ln(\frac{\ln n}{1})^{\ln(n)}}{\ln(n)} = \lim_{n\to\infty} \frac{\ln n \ln(\ln(n))}{\ln(n)} = \lim_{n\to\infty} \ln(\ln(n)) = \infty > 1$ we conclude that $\sum_{n=1}^{\infty} (\frac{1}{\ln(n)})^{\ln(n)}$ converges.

□

Theorem 8.20. *Comparison Remainder Theorem. Let $|a_n| \leq b_n$ for all $n \in \mathbb{N}$ where $\sum_{n=1}^{\infty} b_n$ converges. Let $s_n = \sum_{i=1}^{n} a_i$ for each $n \in \mathbb{N}$ and let $\sum_{n=1}^{\infty} a_n = s$. Then $|s - s_n| \leq \sum_{i=n+1}^{\infty} b_i$.*

Proof. Let $n \in \mathbb{N}$. Since $|\sum_{i=n}^{k} a_i| \leq \sum_{i=n}^{k} |a_i| \leq \sum_{i=n}^{k} b_i$ for each natural number $k > n$, it follows that $|s - s_n| = |\sum_{i=n+1}^{\infty} a_i| \leq \sum_{i=n+1}^{\infty} b_i$ by the Comparison Theorem. □

An interesting property of series convergence relating to absolute convergence relates to the idea of adding a rearrangement of the terms of a sequence. Absolutely convergent series are series so that if the series terms are rearranged in a different order then the sum remains unchanged when the terms are added. For conditionally convergent series, however, the order in which the terms is added is critical, and you can get any sum you wish by appropriately rearranging such a sequence.

Definition 63

Let $g : \mathbb{N} \to \mathbb{N}$ be a one to one and onto function, and let $\{a_n\}$ be a sequence. Then $\{a_{g(n)}\}$ is a *rearrangement* of $\{a_n\}$.

Theorem 8.21. *Let $a_n \geq 0$ for each $n \in \mathbb{N}$ and let $\sum_{n=1}^{\infty} a_n = L$. Let S be a finite subset of the natural numbers. Then $\sum_{n \in S} a_n \leq L$.*

Proof. Let $s_n = \sum_{i=1}^{n} a_i$ denote the nth partial sum of $\{a_i\}$. Then $\{s_n\}$ is increasing and converges to its least upper bound L. Let $M = \max(S)$. Then $\sum_{n \in S} a_n \leq s_M \leq L$ since the terms added in s_M are all non-negative and include every summand in $\sum_{n \in S} a_n$.

□

Theorem 8.22. *Let $\sum_{n=1}^{\infty} a_n$ be an absolutely convergent series with sum s. Let $g : \mathbb{N} \to \mathbb{N}$ be a one to one and onto function, and let $\{b_n\} = \{a_{g(n)}\}$ be a rearrangement of $\{a_n\}$. Then $\sum_{n=1}^{\infty} b_n = s$.*

Proof. Let $\epsilon > 0$. Let $L = \sum_{n=1}^{\infty} |a_n|$. Choose a natural number k so that if $n \geq k$ then $L - \sum_{i=1}^{n-1} |a_i| < \epsilon$, so $\sum_{i=n}^{\infty} |a_i| < \epsilon$. Let $N = \max\{g(1), ..., g(k)\}$. Then for all $1 \leq i \leq k$ it is true that $a_i = g(j)$ for some $1 \leq j \leq N$.

Let $m > N$. If we look at the difference $\sum_{i=1}^{m} a_i - \sum_{i=1}^{m} b_i$, all terms a_i for $1 \leq i \leq k$ cancel with terms b_j for some $1 \leq j \leq N < m$. The terms remaining are terms of the form $\pm a_i$ for $i > k$. Let S be the set of indices i so that $\pm a_i$ is a summand of $\sum_{i=1}^{m} a_i - \sum_{i=1}^{m} b_i$ for some $i \leq m$. In other words $S = \{i \in \mathbb{N} | i \leq m$ and $a_i \neq b_j$ for any $1 \leq j \leq m$ or $a_i = b_j$ for some $1 \leq j \leq m$ so that $b_j \neq a_i$ for any $1 \leq i \leq m\}$. Then $\sum_{i \in S} |a_i| \leq \sum_{i=k+1}^{\infty} |a_i| < \epsilon$ by Theorem 8.21. Thus, $|\sum_{i=1}^{m} a_i - \sum_{i=1}^{m} b_i| \leq \sum_{i \in S} |a_i| < \epsilon$. From this we conclude that the sequence $\{\sum_{i=1}^{m} a_i - \sum_{i=1}^{m} b_i\} \to 0$, so $\sum_{n=1}^{\infty} b_n = s$.

□

Theorem 8.23. *Let $\sum_{n=1}^{\infty} a_n = L \in \mathbb{R}$. Then $\sum_{n=1}^{\infty} a_n$ is conditionally convergent if and only if there are subsequences $\{a_{p(i)}\}$ consisting of all non-negative terms of $\{a_n\}$ and $\{a_{n(i)}\}$ consisting of all negative terms of $\{a_n\}$ having the property that $\sum_{i=1}^{\infty} a_{p(i)} = \infty$ and $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$.*

Proof. First, assume that $\sum_{n=1}^{\infty} a_n$ is conditionally convergent. Note that if there are only finitely many negative terms of $\{a_n\}$ then if we set $S = \{i \in \mathbb{N} | a_i < 0\}$ we would have, for any integer $m > \max(S)$, that $\sum_{i=1}^{m} |a_i| = \sum_{i=1}^{m} a_i + 2\sum_{i \in S} |a_i|$, which means that $\sum_{i=1}^{\infty} |a_i| = L + 2\sum_{i \in S} |a_i|$, so $\sum_{n=1}^{\infty} a_n$ would be absolutely convergent. Similarly, if there were only finitely many indices i so that $a_i \geq 0$ then the series would be absolutely convergent. Thus, there are subsequences $\{a_{p(i)}\}$ and $\{a_{n(i)}\}$ of $\{a_n\}$ so that $p(i)$ and $n(i)$ are the ith non-negative terms and negative terms, respectively, of $\{a_n\}$.

Next, suppose that $\sum_{i=1}^{\infty} a_{p(i)} = P$ and $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$. Let $M < 0$. Then we can choose k so that if $i \geq k$ then $\sum_{i=1}^{\infty} a_{n(i)} < M - P$. Let $m \geq n(k)$. Set $S = \{i \in \mathbb{N} | p(i) \leq m\}$. Let $T = \{i \in \mathbb{N} | n(i) \leq m\}$. Then $\sum_{i=1}^{m} a_i = \sum_{i \in S} a_{p(i)} + \sum_{i \in T} a_{n(i)}$. We know that $\sum_{i \in S} a_{p(i)} < P$ by Theorem 8.21, and we know that $\sum_{i \in T} a_{n(i)}$ consists of a sum of negative terms including all $a_{n(i)}$ with $i \leq k$, which means that $\sum_{i \in T} a_{n(i)} < M - P$. Thus, $\sum_{i=1}^{m} a_i < M$, so $\sum_{i=1}^{\infty} a_i = -\infty$, which contradicts $\sum_{n=1}^{\infty} a_n = L$. It follows, similarly, that it is impossible for $\sum_{i=1}^{\infty} a_{n(i)} = N$ and $\sum_{i=1}^{\infty} a_{p(i)} = \infty$.

Suppose that $\sum_{i=1}^{\infty} a_{p(i)} = P$ and $\sum_{i=1}^{\infty} a_{n(i)} = N$. Then $\sum_{i=1}^{\infty} |a_{p(i)}| = P$ and $\sum_{i=1}^{\infty} |a_{n(i)}| = -N$. Then for any natural number m it follows that if we set $S = \{i \in \mathbb{N} | p(i) \leq m\}$ and $T = \{i \in \mathbb{N} | n(i) \leq m\}$ then $\sum_{i=1}^{m} |a_i| = \sum_{i \in S} |a_{p(i)}| + \sum_{i \in T} |a_{n(i)}| < P - N$. Hence, $\sum_{i=1}^{\infty} |a_i|$ converges to some value less than or equal to $P - N$.

We conclude that if $\sum_{n=1}^{\infty} a_n$ is conditionally convergent then $\sum_{i=1}^{\infty} a_{p(i)} = \infty$ and $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$.

Finally assume that $\sum_{i=1}^{\infty} a_{p(i)} = \infty$ and $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$. Then given any number $M > 0$ we can find a natural number k so that $\sum_{i=1}^{k} a_{p(i)} > M$, so $\sum_{i=1}^{p(k)} |a_i| \geq \sum_{i=1}^{k} a_{p(i)} > M$, which means that $\sum_{i=1}^{\infty} |a_i| = \infty$ so $\sum_{n=1}^{\infty} a_n$ is not absolutely convergent (and must be conditionally convergent).

□

Theorem 8.24. *Let $\sum_{n=1}^{\infty} a_n$ be a conditionally convergent series. Then if $L \in \mathbb{R}$ or $L = \pm\infty$ then there is one to one and onto function $g : \mathbb{N} \to \mathbb{N}$ so that $\sum_{i=1}^{\infty} a_{g(i)} = L$.*

Proof. By Theorem 8.23, we can find sequences $\{a_{p(i)}\}$ consisting of all positive terms of $\{a_n\}$ and $\{a_{n(i)}\}$ consisting of all negative terms of $\{a_n\}$ having the property that $\sum_{i=1}^{\infty} a_{p(i)} = \infty$ and $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$.

First, assume that $L = \infty$. Let m_1 be the first integer so that $\sum_{i=1}^{m_1} a_{p(i)} > 1$. Then let m_2 be the first positive integer exceeding m_1 so that $\sum_{i=1}^{m_2} a_{p(i)} + a_{n(1)} > 2$. Inductively, choose m_{k+1} to be the first positive integer exceeding m_k so that $\sum_{i=1}^{m_{k+1}} a_{p(i)} + \sum_{i=1}^{m_k} a_{n(i)} > k + 1$. We can always make such a choice since $\sum_{i=1}^{\infty} a_{p(i)} = \infty$. We then define rearrangement $\{b_i\} = \{a_{p(1)}, a_{p(2)}, ..., a_{p(m_1)},\ a_{n(1)}, a_{p(m_1+1)}, a_{p(m_1+2)}, ..., a_{p(m_2)}\ a_{n(2)}, a_{p(m_2+1)}, ...\}$. Note that for any positive integer s it is the case that if $m_s + s \leq t$ then $\sum_{i=1}^{t} b_i \geq s$. Thus, $\sum_{i=1}^{t} b_i = \infty$. Similarly, we can find a rearrangement $\{b_i\}$ whose sum is $-\infty$.

Next, let $L \in \mathbb{R}$. We make the rearrangement choices as follows. We set $b_1 = a_{p(1)}$. If we have chosen the first k members of the rearrangement then if $\sum_{i=1}^{k} b_i \leq L$ we choose b_{k+1} to be $a_{p(t)}$, where t is the first index so that $a_{p(t)} \notin \{b_1, b_2, ..., b_k\}$. If $\sum_{i=1}^{k} b_i > L$ we choose b_{k+1} to be $a_{n(s)}$, where s is the first index so that $a_{n(s)} \notin \{b_1, b_2, ..., b_k\}$. Thus, the partial sums of $\{b_i\}$ are non-decreasing until they exceed L, then non-increasing until the precede L and then non-decreasing until they exceed L again, and so forth. Let $\epsilon > 0$. Since $\sum_{n=1}^{\infty} a_n$ is convergent, we can find an integer m so that if $i \geq m$ then $|a_i| < \epsilon$. We will assume that $\sum_{i=1}^{m} b_i \leq L$ (the case where $\sum_{i=1}^{m} b_i > L$ is similar). Since $\sum_{i=1}^{\infty} a_{p(i)} = \infty$ there will be a first $j_1 > m$ so that $\sum_{i=1}^{j_1} b_i > L$. Then since $\sum_{i=1}^{\infty} a_{n(i)} = -\infty$ there is a first $j_2 > j_1$ so that $\sum_{i=1}^{j_2} b_i \leq L$. For all $j_1 \leq j \leq j_2$ it follows that $L + b_{j_2} \leq s_j \leq L + b_{j_1}$ (where $s_j = \sum_{i=1}^{j} b_i$). Likewise, if j_3 is the first integer so that $j_3 > j_2$ and $\sum_{i=1}^{j_3} b_i > L$ then for all $j_2 \leq j \leq j_3$ it follows that s_j is equal to or between $L + b_{j_2}$ and $L + b_{j_3}$ and so forth. Thus, since each $|b_{j_i}| < \epsilon$ it follows that $|L - s_j| < \epsilon$ for all $j \geq j_1$. Hence, $\sum_{i=1}^{\infty} b_i = L$.

□

The following weaker form of Stirling's Formula is often helpful in using the Log Test to determine the convergence or divergence of series that involve a factorial:

Theorem 8.25. *Weaker form of Stirling's Formula. For each $C > 1$ there is a $k \in \mathbb{N}$ so that if n is a natural number larger than k then $n\ln(n) - n < \ln(n!) < n\ln(n) - n + C\ln(n)$.*

Proof. First, note $\ln(n!) = \ln(1) + \ln(2) + ... + \ln(n) = \sum_{i=2}^{n} \ln(i)$. For any natural number n we note that since $\ln(x)$ is an increasing function, it must be the case that $\int_{n-1}^{n} \ln(x)dx < \ln(n) < \int_{n}^{n+1} \ln(x)dx$. Thus, $\sum_{i=2}^{n} \int_{i-1}^{i} \ln(x)dx < \sum_{i=2}^{n} \ln(i) < \sum_{i=2}^{n} \int_{i}^{i+1} \ln(x)dx$, so $\int_{1}^{n} \ln(x)dx \leq \ln(n!) < \int_{2}^{n+1} \ln(x)dx < \int_{1}^{n+1} \ln(x)dx$. Evaluating these integrals, we get $n\ln(n) - n + 1 < \ln(n!) < (n+1)\ln(n+1) - (n+1) + 1$. Thus, $(n\ln(n) - n) < \ln(n!) < (n+1)\ln(n+1) - n$, which means that $0 < \ln(n!) - (n\ln(n) - n) < n\ln(n+1) - n\ln(n) + \ln(n+1) + (n - n)$, so $0 < \ln(n!) - (n\ln(n) - n) < n\ln(1 + \frac{1}{n}) + \ln(n+1)$. Notice $n\ln(1 + \frac{1}{n}) = \ln(1 + \frac{1}{n})^n$ which approaches $\ln(e) = 1$ as n approaches infinity. Also, $\ln(n+1) - \ln(n) = \ln(1 + \frac{1}{n})$

approaches zero. Thus, if $\epsilon = C - 1 > 0$ then we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $n \ln(1 + \frac{1}{n}) < 2 < \frac{\epsilon}{2} \ln(n)$ and $\ln(n + 1) < (1 + \frac{\epsilon}{2}) \ln(n)$ which means that $n \ln(1 + \frac{1}{n}) + \ln(n + 1) < (1 + \epsilon) \ln(n) = C \ln(n)$.

□

We will refrain from giving a list of hints for the exercises in the remaining sections. It is hoped that readers have a fairly good idea where to start looking at definitions and theorems relating to a problem at this point, having practiced the process in earlier sections.

Exercises:

Exercise 8.1. *Prove that* $\sum_{n=1}^{\infty} \frac{1}{n^p}$ *converges if and only if* $p > 1$.

Exercise 8.2. *Let* $a_n = \frac{n^n}{n!}$. *Prove that* $\lim_{x\to\infty} \frac{a_{n+1}}{a_n} = e$.

Exercise 8.3. *Determine whether each series converges or diverges, with justification.*

(a) $\sum_{n=1}^{\infty} \frac{n!}{n^n}$

(b) $\sum_{n=1}^{\infty} \frac{\cos(n)}{n^2}$

(c) $\sum_{n=1}^{\infty} (-1)^n (1 - \frac{2}{n})^n$

(d) $\sum_{n=1}^{\infty} \frac{\ln(n)}{n^2}$.

(e) $\sum_{n=2}^{\infty} \frac{1}{(\ln n)^{\ln n}}$

Exercise 8.4. *Give an example, with justification, of a series* $\sum_{n=1}^{\infty} (-1)^n a_n$ *which is divergent, where* $a_n > 0$ *for all* n *and* $\{a_n\} \to 0$.

Exercise 8.5. *Give an example of a series* $\sum_{n=1}^{\infty} (-1)^n a_n$ *which is convergent and* $\sum_{n=1}^{\infty} (-1)^n b_n$ *which is divergent so that* $\{\frac{a_n}{b_n}\} \to L \neq 0$.

Exercise 8.6. *Let* $|f^{(n)}(x)| \leq M_n$ *on the interval* $[a, b]$ *for all* $n \in \mathbb{N}$, *and let* $c \in [a, b]$. *Then* $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)(x-c)^n}{n!}$ *for all* $x \in [a, b]$ *if* $\lim_{n\to\infty} \frac{M_n(b-a)^n}{(n)!} = 0$.

Exercise 8.7. *Give an example of series* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *which both converge, so that* $\sum_{n=1}^{\infty} a_n b_n$ *diverges.*

Exercise 8.8. *Prove that if* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *converge absolutely then* $\sum_{n=1}^{\infty} a_n b_n$ *must also converge.*

Exercise 8.9. *Prove that if* $\limsup\{|c_n|^{\frac{1}{n}}\} \to L < 1$ *then* $\lim_{n\to\infty} c_n = 0$*.*

Solutions to Exercises in Chapter 8.

Solution to Exercise 8.1. *Prove that* $\displaystyle\sum_{n=1}^{\infty}\frac{1}{n^p}$ *converges if and only if* $p > 1$.

Proof. Using the integral test we have $\displaystyle\int_1^{\infty}\frac{1}{x^p}dx = \lim_{b\to\infty}\ln(b) = \infty$ if $p = 1$ and otherwise $\displaystyle\int_1^{\infty}\frac{1}{x^p}dx = \lim_{b\to\infty}\frac{-1}{(p-1)x^{p-1}} = \infty$ if $p < 1$ or $\displaystyle\frac{1}{p-1}$ if $p > 1$. □

Solution to Exercise 8.2. *Let* $a_n = \dfrac{n^n}{n!}$. *Prove that* $\displaystyle\lim_{x\to\infty}\frac{a_{n+1}}{a_n} = e$.

Proof. We take $\displaystyle\lim_{n\to\infty}\frac{(n+1)^{n+1}}{(n+1)!}\frac{n!}{n^n} = \lim_{n\to\infty}\frac{(n+1)^n}{n^n} = \lim_{n\to\infty}(1+\frac{1}{n})^n = e$. □

Solution to Exercise 8.3. *Determine whether each series converges or diverges, with justification.*

(a) $\displaystyle\sum_{n=1}^{\infty}\frac{n!}{n^n}$

(b) $\displaystyle\sum_{n=1}^{\infty}\frac{\cos(n)}{n^2}$

(c) $\displaystyle\sum_{n=1}^{\infty}(-1)^n(1-\frac{2}{n})^n$

(d) $\displaystyle\sum_{n=1}^{\infty}\frac{\ln(n)}{n^2}$.

(e) $\displaystyle\sum_{n=2}^{\infty}\frac{1}{(\ln n)^{\ln n}}$

Proof. (a) Using the Ratio Test and the preceding exercise we have $\displaystyle\lim_{n\to\infty}\frac{(n+1)!}{(n+1)^{n+1}}\frac{n^n}{n!} = \frac{1}{e} < 1$. Thus, the series converges. We could also have used the Comparison Test with $\dfrac{1}{n^2}$, the Log Test, or even the Root Test (though the Root Test is a bit trickier).

(b) Using the Comparison Test we have that $\dfrac{|\cos(n)|}{n^2} \leq \dfrac{1}{n^2}$. Since $\displaystyle\sum_{n=1}^{\infty}\frac{1}{n^2}$ converges by Exercise 8.1, it follows that $\displaystyle\sum_{n=1}^{\infty}\frac{\cos(n)}{n^2}$ converges absolutely and therefore converges.

(c) We can use the Divergence or Log Tests. If we set $u = \dfrac{n}{-2}$ then $n = -2u$, so $\{(1-\frac{2}{n})^n\} = \{((1+\frac{1}{u})^u)^{-2}\} \to \frac{1}{e^2} \neq 0$, which means that $(-1)^n(1-\frac{2}{n})^n \not\to 0$, and hence $\displaystyle\sum_{n=1}^{\infty}(-1)^n(1-\frac{2}{n})^n$ diverges by the Divergence Test.

(d) We can use the Comparison or Log Test. For the Comparison Test, we note that $\sqrt{x} > \ln(x)$ for sufficiently large x (actually, for all $x > 0$, but we do not need that). One way to see this is using L'Hospital's Rule to find $\lim\limits_{x\to\infty} \dfrac{\ln(x)}{\sqrt{x}} = \lim\limits_{x\to\infty} \dfrac{\frac{1}{x}}{\frac{1}{2\sqrt{x}}} = \lim\limits_{x\to\infty} \dfrac{2}{\sqrt{x}} = 0$. Thus, for some M if $x \geq M$ then $\dfrac{\ln(x)}{\sqrt{x}} < 1$ which means that $\ln(x) < \sqrt{x}$. Once this has been established we can use the Comparison Test by noting that for sufficiently large values of n it is true that $\dfrac{\ln(n)}{n^2} < \dfrac{\sqrt{n}}{n^2} = \dfrac{1}{n^{\frac{3}{2}}}$. We know that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^{\frac{3}{2}}}$ converges by Exercise 8.1, and so $\displaystyle\sum_{n=1}^{\infty} \frac{\ln(n)}{n^2}$ converges by the Comparison Test.

The Log Test may be shorter. You take $\lim\limits_{n\to\infty} \dfrac{\ln(\frac{n^2}{\ln(n)})}{\ln(n)} = \lim\limits_{n\to\infty} \dfrac{2\ln(n) - \ln(\ln(n))}{\ln(n)} = \lim\limits_{n\to\infty} 2 - \dfrac{\ln(\ln(n))}{\ln(n)} = 2 > 1$ since $\lim\limits_{n\to\infty} \dfrac{\ln(\ln(n))}{\ln(n)} = \lim\limits_{n\to\infty} \dfrac{n}{1}\dfrac{1}{n\ln(n)} = \lim\limits_{n\to\infty} \dfrac{1}{\ln(n)} = 0$ by L'Hospital's Rule.

(e) Using the log test we have $\lim\limits_{n\to\infty} \dfrac{\ln(\ln(n))^{\ln(n)}}{\ln(n)} = \lim\limits_{n\to\infty} \ln(\ln(n)) = \infty$, so the series converges.

□

Solution to Exercise 8.4. *Give an example, with justification, of a series* $\displaystyle\sum_{n=1}^{\infty}(-1)^n a_n$ *which is divergent, where* $a_n > 0$ *for all* n *and* $\{a_n\} \to 0$*.*

Proof. We could use $a_n = \dfrac{1}{n}$ if n is odd, and $a_n = \dfrac{1}{2^n}$ when n is even. We know that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ by Exercise 8.1 since the partial sums s_n of this series are increasing and thus only diverge if $\{s_n\}$ is not bounded above and hence becomes and remains larger than any given number for sufficiently large n. Thus, $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2n} = \infty$ and since $\dfrac{1}{2n} < \dfrac{1}{2n-1}$ it follows that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2n-1} = \infty$ by the Comparison Test. Thus, $-1 - \dfrac{1}{3} - \dfrac{1}{5} - ... \to -\infty$ whereas $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2^{2n}} \to \frac{1}{3}$. Hence, given any $M < 0$ we can find an integer k so that $\displaystyle\sum_{n=1}^{k} \frac{-1}{2k-1} =< M - \frac{1}{3}$ which means that if $N = 2k+1$ then if $j \geq N$ we have that $\displaystyle\sum_{n=1}^{j}(-1)^n a_n \leq \sum_{n=1}^{k} \frac{-1}{2k-1} + \sum_{n=1}^{N} \frac{1}{2^{2n}} < M - \frac{1}{3} + \frac{1}{3} = M$. Hence, $\displaystyle\sum_{n=1}^{j}(-1)^n a_n = -\infty$. □

Solution to Exercise 8.5. *Give an example of a series* $\sum_{n=1}^{\infty}(-1)^n a_n$ *which is convergent and* $\sum_{n=1}^{\infty}(-1)^n b_n$ *which is divergent so that* $\{\frac{a_n}{b_n}\} \to L \neq 0$.

Proof. Let $\{a_n\} = \frac{1}{\sqrt{n}}$ and let $b_n = \frac{1}{\sqrt{n}} + \frac{(-1)^{n+1}}{n}$. Then $\frac{b_n}{a_n} = 1 + \frac{(-1)^{n+1}}{\sqrt{n}}$ which converges to $1 \neq 0$, so $\frac{a_n}{b_n} \to 1$. By the alternating series test, $\sum_{n=1}^{\infty}(-1)^n a_n$ converges, but $\sum_{n=1}^{\infty}(-1)^n b_n = \sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}} - \frac{1}{n}$. We know that the partial sums of $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ are bounded and that the partial sums of $\sum_{n=1}^{\infty} \frac{1}{n}$ diverge to infinity, from which we can conclude that $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}} - \frac{1}{n} = -\infty$. □

Solution to Exercise 8.6. *Let* $|f^{(n)}(x)| \leq M_n$ *on the interval* $[a, b]$ *for all* $n \in \mathbb{N}$, *and let* $c \in [a, b]$. *Then* $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)(x-c)^n}{n!}$ *for all* $x \in [a, b]$ *if* $\lim_{n\to\infty} \frac{M_n(b-a)^n}{(n)!} = 0$.

Proof. This follows immediately from the Squeeze Theorem and the second form of Taylor's Theorem, which states that $f(x) = \sum_{n=0}^{n} \frac{f^{(i)}(c)(x-c)^i}{i!} + \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$. Since we know that $0 \leq |\frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!}$, by the Squeeze Theorem, we can conclude that $\lim_{n\to\infty} \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!} = 0$, which means $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)(x-c)^n}{n!}$. □

Solution to Exercise 8.7. *Give an example of series* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *which both converge, so that* $\sum_{n=1}^{\infty} a_n b_n$ *diverges.*

Proof. Take $a_n = b_n = \frac{(-1)^n}{\sqrt{n}}$. Their product $\frac{1}{n}$ is a sequence whose sum diverges by Exercise 8.1, and the original series converge by the Alternating Series Test. □

Solution to Exercise 8.8. *Prove that if* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *converge absolutely then* $\sum_{n=1}^{\infty} a_n b_n$ *must also converge.*

Proof. Since $\sum_{n=1}^{\infty} b_n$ converges we know that $\{b_n\} \to 0$ be the Divergence Test, so for some k if $n \geq k$ then $|b_n| < 1$. Thus, $|a_n b_n| < |a_n|$. Since $\sum_{n=1}^{\infty} |a_n|$ converges, by the Comparison Test we know $\sum_{n=1}^{\infty} |a_n b_n|$ converges, which implies that $\sum_{n=1}^{\infty} a_n b_n$ converges. □

Solution to Exercise 8.9. *Prove that if* $\limsup\{|c_n|^{\frac{1}{n}}\} \to L < 1$ *then* $\lim_{n\to\infty} c_n = 0$.

Proof. Using the Root test we see that $\sum_{n=1}^{\infty} c_n$ converges, so $\{c_n\} \to 0$ by the Divergence Test. □

Chapter 9

Sequences of Functions

Definition 64

Let $f_n : D \to \mathbb{R}$ for each $n \in \mathbb{N}$. We say $\{f_n\} \to f$ for a function $f : D \to \mathbb{R}$ (or converges *pointwise* to f) if, for each $x \in D$, $\{f_n(x)\} \to f(x)$. We say that $\{f_n\} \to f$ *uniformly* on D if for every $\epsilon > 0$ there is a $k \in \mathbb{N}$ so that if $n \geq k$ then $|f_n(x) - f(x)| < \epsilon$ for all $x \in D$.

Most useful properties are not preserved under pointwise convergence of sequences of functions, but many are preserved under uniform convergence.

Theorem 9.1. *Let $f_n : D \to \mathbb{R}$ be continuous at x_0 for each $n \in \mathbb{N}$, and let $\{f_n\} \to f$ uniformly on D. Then f is continuous at x_0.*

Proof. Let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|f_n(x) - f(x)| < \dfrac{\epsilon}{3}$ for all $x \in D$. Choose $\delta > 0$ so that if $|x - x_0| < \delta$ and $x \in D$ then $|f_k(x) - f_k(x_0)| < \dfrac{\epsilon}{3}$. Then if $|x - x_0| < \delta$ and $x \in D$, it follows that $|f(x) - f(x_0)| \leq |f_k(x) - f(x)| + |f_k(x) - f_k(x_0)| + |f_k(x_0) - f(x_0)| < \epsilon$. Thus, f is continuous at x_0.

□

Theorem 9.2. *The Uniform Cauchy Criterion. Let $f_n : D \to \mathbb{R}$ for each $n \in \mathbb{N}$. Then there is a function f so that $\{f_n\} \to f$ uniformly on D if and only if, for every $\epsilon > 0$, there is a $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|f_n(x) - f_m(x)| < \epsilon$ for all $x \in D$.*

Proof. Assume first that for every $\epsilon > 0$ there is a $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|f_n(x) - f_m(x)| < \epsilon$ for all $x \in D$. For each $x \in D$ we note that for every $\epsilon > 0$ there is a $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|f_n(x) - f_m(x)| < \epsilon$. Hence, $\{f_n(x)\}$ is a Cauchy sequence which by previous theorems converges to a point which we define to be $f(x)$.

Let $\epsilon > 0$. Then we can find a $k \in \mathbb{N}$ so that if $n, m \geq k$ then $|f_n(x) - f_m(x)| < \dfrac{\epsilon}{2}$ for all $x \in D$. Thus, $|f_n(x) - f(x)| = \lim\limits_{m \to \infty} |f_n(x) - f_m(x)| \leq \dfrac{\epsilon}{2} < \epsilon$, so $\{f_n\} \to f$ uniformly.

Next, assume that there is a function f so that $\{f_n\} \to f$ uniformly on D and let $\epsilon > 0$. Then choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|f_n(x) - f(x)| < \frac{\epsilon}{2}$. If $n, m \geq k$ then it follows that $|f_n(x) - f_m(x)| \leq |f_n(x) - f(x)| + |f_m(x) - f(x)| < \epsilon$.

□

Theorem 9.3. *Let $f_n : D \to \mathbb{R}$ be integrable on $[a, b]$ for each $n \in \mathbb{N}$, where $\{f_n\}$ converges uniformly to the function $f(x)$ on a closed interval $[a, b]$. Then $\{\int_a^x f_n\} \to \int_a^x f$ uniformly on $[a, b]$. In particular, $\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx$.*

Proof. First, if we let E_i be the set of points at which f_i is discontinuous and E be the set of points at which f is discontinuous then $E \subseteq \bigcup_{i=1}^{\infty} E_i$ by Theorem 9.1, and so $\lambda(E) = 0$ by Theorem 7.62, and thus f is integrable by the Lebesgue Characterization of Riemann Integrability. Let $\epsilon > 0$ and let k be an integer such that if $n \geq k$ then $|f(x) - f_n(x)| < \frac{\epsilon}{3(b-a)}$ for all $x \in [a, b]$. Choose $m \geq k$ and $x \in [a, b]$, and select a partition $P = \{x_0, x_1, x_2, ..., x_n\}$ of $[a, x]$ so that $U(f_m, P) - L(f_m, P) < \frac{\epsilon}{3}$ and $U(f, P) - L(f, P) < \frac{\epsilon}{3}$. Choose any marking $T = \{x_1^*, ..., x_n^*\}$ of P. Then $|S_T(f_m, P) - S_T(f, P)| = |\sum_{i=1}^{n}(f(x_i^*) - f_m(x_i^*))(x_i - x_{i-1})| < \frac{\epsilon}{3(b-a)} \sum_{i=1}^{n} x_i - x_{i-1} = (b-a)(\frac{\epsilon}{3(b-a)}) = \frac{\epsilon}{3}$. Thus, $|\int_a^x f_m - \int_a^x f| \leq |\int_a^x f_m - S_T(f_m, P)| + |S_T(f_m, P) - S_T(f, P)| + |S_T(f, P) - \int_a^x f| < \epsilon$. Thus, $\{\int_a^x f_n\} \to \int_a^x f$ uniformly on $[a, b]$ and $\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx$.

□

Theorem 9.4. *Let f_n be continuously differentiable on $[a, b]$ for each $n \in \mathbb{N}$, and let $\{f_n'\} \to f'$ uniformly on $[a, b]$, and for some $s \in (a, b)$ let $\{f_n(s)\} \to f(s)$. Then $\{f_n(x)\} \to f(x)$ uniformly on $[a, b]$.*

Proof. Note that each $f_n(x) = \int_s^x f_n'(t)dt + f_n(s)$ by the Fundamental Theorem of Calculus, since each f_n' is continuous and therefore integrable. Since $\{f_n'\} \to f'$ uniformly on $[a, b]$, it follows from 9.1 that f' is continuous on $[a, b]$ as well so by the Fundamental Theorem of Calculus $f(x) = \int_s^x f'(t)dt + f(s)$. Thus, by Theorem 9.3 it follows that $\{f_n(x)\} = \{\int_s^x f_n'(t)dt + f_n(s)\} \to \int_s^x f'(t)dt + f(s) = f(x)$ uniformly on $[a, b]$ and therefore $\{f_n(x)\} \to f(x)$ uniformly by Exercise 9.9 . □

Theorem 9.5. *Weierstrass M-test. Let $|f_n(x)| \leq M_n$ on E for each $n \in \mathbb{N}$ and let $\sum_{n=1}^{\infty} M_n$ converge. Then $\sum_{n=1}^{\infty} f_n$ converges absolutely and uniformly on E.*

Proof. Absolute convergence follows directly from the Comparison Test. Let $\epsilon > 0$. Since $\sum_{n=1}^{\infty} M_n$ converges, by the Cauchy Convergence Criterion we can find $k \in \mathbb{N}$ so that if $n > m \geq k$ then $\sum_{i=m+1}^{n} M_i < \epsilon$, so $|\sum_{i=m+1}^{n} f_i(x)| \leq \sum_{i=m+1}^{n} |f_i(x)| <\leq \sum_{i=m+1}^{n} M_i \epsilon$ for all $x \in E$. Hence, $\sum_{n=1}^{\infty} f_n$ converges uniformly on E by the Uniform Cauchy Criterion. □

Definition 65

A series of the form $\sum_{n=0}^{\infty} c_n(x-a)^n$ is called a *power series centered at a.* We say that a function f is *analytic* on an interval (s,t) if for every $x_0 \in (s,t)$ there is an $\epsilon > 0$ so $f(x) = \sum_{n=0}^{\infty} c_n(x-x_0)^n$ for some coefficients c_n for all $x \in (x_0 - \epsilon, x_0 + \epsilon)$. If f is infinitely differentiable on (s,t) then we call $\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)(x-x_0)^n}{n!}$ the *Taylor series* for f centered at x_0. If $x_0 = 0$ we call the series a Maclaurin series.

If a power series $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges only if $x = a$ then we say the *radius of convergence* of the power series is zero. If there is a positive number R so that the series converges on $(a-R, a+R)$ but diverges on $\mathbb{R} \setminus [a-R, a+R]$ then we say the series has radius of convergence of the series is R. If the series converges on $\mathbb{R}$ then we say that the radius of convergence is ∞. The *interval of convergence* of a power series is the set of values for which the power series converges.

We will show that the "interval of convergence" is an interval in theorems that follow.

Theorem 9.6. *Let $\sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series. Then:*

(a) Exactly one of the following is true:

(1) $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges only at the point a and so the radius of convergence R of the series is zero, which is true if and only if $\limsup |c_n|^{\frac{1}{n}} = \infty$.

(2) $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely for all real x and the radius of convergence R of the series is infinity, which is true if and only if $\limsup |c_n|^{\frac{1}{n}} = 0$.

(3) There is a positive number R so that $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely on $(a-R, a+R)$ and diverges for all $x \in \mathbb{R} \setminus [a-R, a+R]$, meaning that the radius of convergence of the series is R, which is true if and only if $\limsup |c_n|^{\frac{1}{n}} = \frac{1}{R}$.

(b) If $r > 0$ then $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely on $(a-r, a+r)$ if and only if $r \limsup |c_n|^{\frac{1}{n}} \leq 1$.

(c) If $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges on an interval $(a-R, a+R)$ and $[c,d] \subset (a-R, a+R)$ then $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges uniformly on $[c,d]$.

Proof. We use the Root Test. First, note that $\sum_{n=0}^{\infty} c_n(x-a)^n = c_0$ if $x = a$ regardless of choice of $\{c_n\}$.

(a) If $\{|c_n|^{\frac{1}{n}}\}$ is unbounded (meaning $\limsup |c_n|^{\frac{1}{n}} = \infty$) if and only if $|c_n|^{\frac{1}{n}}|x-a|$ is unbounded if $x \neq a$, so $\limsup |c_n|^{\frac{1}{n}}|x-a| = \infty$ if and only if $\limsup |c_n|^{\frac{1}{n}} = \infty$ unless $x = a$. In this case, $\sum_{n=0}^{\infty} c_n(x-a)^n$ diverges, so the radius of convergence is zero.

If $\{c_n\}$ is bounded then $\limsup |c_n|^{\frac{1}{n}}$ is a non-negative number. If $x \neq a$ then $\limsup |c_n|^{\frac{1}{n}} = 0$ if and only if $\limsup |c_n|^{\frac{1}{n}}|x-a| = 0$ by Exercise 9.1, in which case $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely for each $x \in \mathbb{R}$ by the Root Test, so the radius of convergence of the series is infinity in this case.

If $\limsup |c_n|^{\frac{1}{n}}$ is positive then let $R = (\limsup |c_n|^{\frac{1}{n}})^{-1}$. Then $\limsup |c_n|^{\frac{1}{n}}|x-a| = |x-a| \limsup |c_n|^{\frac{1}{n}}$ by Exercise 9.1. Since $|x-a| \limsup |c_n|^{\frac{1}{n}} < 1$ if $|x-a| < R$ and $|x-a| \limsup |c_n|^{\frac{1}{n}} > 1$ if $|x-a| > R$, we know that $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely if $|x-a| < R$ and diverges if $|x-a| > R$ by the Root Test. Hence, the radius of convergence of the series is R in this case.

(b) As noted above, if $r \limsup |c_n|^{\frac{1}{n}} \leq 1$ for some positive number r then if $|x-a| < r$ we know that $|x-a| \limsup |c_n|^{\frac{1}{n}} < 1$, so the series $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges absolutely for all $x \in (a-r, a+r)$ by the Root Test. Conversely, if $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges (absolutely) for all $x \in (a-r, a+r)$ then that means either the radius of convergence of the series is infinite and thus $\limsup |c_n|^{\frac{1}{n}} = 0$ by part (a) and hence $r \limsup |c_n|^{\frac{1}{n}} = 0 < 1$, or the radius of convergence is equal to some positive number $R \geq r$, in which case $R \limsup |c_n|^{\frac{1}{n}} = 1$ by part (a). Since $r \leq R$ it follows that $R \limsup |c_n|^{\frac{1}{n}} \leq 1$.

(c) If $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges on $(a-R, a+R)$ then the radius of convergence of the series is at least R so $w = \limsup |c_n|^{\frac{1}{n}} \leq \frac{1}{R}$. We can find a positive number $u < R$ so that $[c,d] \subset (a-u, a+u)$ and $\limsup |c_n|^{\frac{1}{n}} |x-a| \leq uw < 1$ and $|c_n(x-a)^n| < (uw)^n$ for all $x \in [c,d]$. Hence, if we set $M_n = (uw)^n$ then $\sum_{n=1}^{\infty} M_n$ is a geometric series which converges, and so $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges uniformly on $[c,d]$ by the Weierstrass M-test.

□

Theorem 9.7. *Let $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ be a power series and let $R > 0$. Then $\sum_{n=0}^{\infty} c_n(x-a)^n$ converges on the interval $(a-R, a+R)$ if and only if $\sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$ converges on the interval $(a-R, a+R)$, which is true if and only if $\sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1}$ converges on the interval $(a-R, a+R)$.*

Proof. Note that $\limsup n^{\frac{1}{n}} |c_n|^{\frac{1}{n}} = \limsup |c_n|^{\frac{1}{n}} = \limsup (\frac{1}{n+1})^{\frac{1}{n}} |c_n|^{\frac{1}{n}}$ since $\{n^{\frac{1}{n}}\} \to 1$ by exercises 9.4 and 9.1. Since each of $\sum_{n=0}^{\infty} c_n(x-a)^n$, $\sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$, and $\sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1}$ converge on the interval $(a-R, a+R)$ if and only if $R \limsup |c_n|^{\frac{1}{n}} \leq 1$, the result follows. □

Theorem 9.8. *Let $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ for all $x \in (a-R, a+R)$. Then $f'(x) = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$ and the derivative of $\sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1}$ is $f(x)$ for all $x \in (a-R, a+R)$.*

Proof. For each $x \in (a-R, a+R)$ we can choose $[c,d]$ so that $a, x \in (c,d) \subset [c,d] \subset (a-R, a+R)$. By Theorem 9.6, f converges uniformly on $[c,d]$ and by Theorem 9.3, $\{\int_a^x \sum_{i=0}^{n} c_i(t-a)^i dt\} = \{\sum_{i=0}^{n} \frac{c_i}{i+1}(x-a)^{i+1}\} \to \int_a^x f(t)dt$ uniformly. By the first form of the Fundamental Theorem of Calculus, $f(x) = (\int_a^x f(t)dt)' = (\sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1})'$. By Theorem 9.7 and Theorem 9.6, $g(x) = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$ converges uniformly on $[c,d]$. Hence, as before, by Theorem 9.3 we know that $\int_a^x g(t)dt = \sum_{n=1}^{\infty} c_n(x-a)^n = f(x) - c_0$. Thus, $f'(x) = (f(x) - c_0)' = g(x)$. □

Theorem 9.9. *Let $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ for all $x \in (a-R, a+R)$. Then $c_n = \dfrac{f^{(n)}(a)}{n!}$ for each $n \in \mathbb{N}$.*

Proof. By Theorem 9.8, the function f is infinitely differentiable on $(a-R, a+R)$, $f'(x) = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1}$, $f''(x) = \sum_{n=2}^{\infty} n(n-1)c_n(x-a)^{n-2}$ and inductively, $f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)(n-2)...(n-k+1)c_n(x-a)^{n-k}$ for each $k \in \mathbb{N}$. Thus, $f^{(k)}(a) = k!c_k + 0 + 0 + ...$, so $c_k = \dfrac{f^{(k)}(a)}{k!}$. □

Theorem 9.10. *Let f be an infinitely differentiable function on an interval $I = [a, b]$ so that, for some $M > 0$, $|f^{(n)}(x)| \leq M^n$ for all $x \in I$ and each $n \in \mathbb{N}$. Then $f(x) = \sum_{n=0}^{\infty} \dfrac{f^{(n)}(a)(x-a)^n}{n!}$ converges uniformly on I.*

Proof. By Taylor's theorem, $f(x) = \sum_{i=0}^{n} \dfrac{f^{(i)}(a)(x-a)^i}{i!} + \dfrac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$ for some point $c \in [a, x]$, which means that $|f(x) - \sum_{i=0}^{n} \dfrac{f^{(i)}(a)(x-a)^i}{i!}| \leq \dfrac{M^{n+1}(x-a)^{n+1}}{(n+1)!}$ which converges to zero. Hence, $\sum_{n=0}^{\infty} \dfrac{f^{(n)}(a)(x-a)^n}{n!}$ converges uniformly to $f(x)$.

□

Theorem 9.11. *The following are Maclaurin series converging to the functions listed on the intervals listed. In each case, we refer to the nth degree Taylor polynomial centered at zero as $T_n(x) \sum_{i=1}^{n} \dfrac{f^{(i)}(0)x^i}{i!}$, and the Taylor Remainder after the nth power term as $R_n = \dfrac{1}{n!} \int_0^x f^{(n+1)}(t)(x-t)^{n+1} dt$. We also uses $T(x)$ to refer to $\sum_{i=1}^{\infty} \dfrac{f^{(i)}(0)x^i}{i!}$, the Maclaurin series for f.*

(a) $\dfrac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + ...$ *on* $(-1, 1)$.

(b) $\dfrac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n = 1 - x + x^2 - x^3 + ...$ *on* $(-1, 1)$.

(c) $\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \dfrac{x^n}{n} = x - \dfrac{x^2}{2} + \dfrac{x^3}{3} - ...$ *on* $(-1, 1]$.

(d) $\tan^{-1}(x) = \sum_{n=0}^{\infty} (-1)^n \dfrac{x^{2n+1}}{2n+1} = x - \dfrac{x^3}{3} + \dfrac{x^5}{5} - ...$ *on* $[-1, 1]$.

(e) $(1+x)^\alpha = \sum_{n=0}^{\infty} \frac{\alpha(\alpha-1)(\alpha-2)...(\alpha-n+1)}{n!}x^n = 1+\alpha x+\frac{\alpha(\alpha-1)}{2!}x^2+\frac{\alpha(\alpha-1)(\alpha-2)}{3!}x^3+$ *... on* $(-1,1)$*, where* $\alpha \in \mathbb{R}$*. We refer to* $\frac{\alpha(\alpha-1)(\alpha-2)...(\alpha-n+1)}{n!}$ *as* $\binom{\alpha}{n}$ *so we can write* $(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n$*. This Maclaurin series is referred to as the Binomial Series.*

(f) $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$ *on* $(-\infty, \infty)$

(g) $\sin(x) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ...$ *on* $(-\infty, \infty)$

(h) $\cos(x) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ...$ *on* $(-\infty, \infty)$

(i) $\sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + ...$ *on* $(-\infty, \infty)$

(j) $\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + ...$ *on* $(-\infty, \infty)$

Proof. We demonstrate a useful preliminary observation that lets us extend Taylor series convergent at their end points to corresponding function values:

If $T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)(x-a)^n}{n!} = f(x)$ for a C^∞ function f on an interval $(a-\epsilon, a+\epsilon)$ then if f is continuous at $a+\epsilon$ and T converges uniformly on some interval $[c, a+\epsilon]$ then and T is continuous at $a+\epsilon$ by Theorem 9.1 and therefore $T(a+\epsilon) = f(a+\epsilon)$ because $f(a+\epsilon) = \lim_{x\to a+\epsilon^-} f(x) = \lim_{x\to a+\epsilon^-} T(x) = T(a+\epsilon)$. Similarly, if T converges uniformly on some interval $[a-\epsilon, c]$ and f is continuous at $a-\epsilon$ then $f(a-\epsilon) = T(a-\epsilon)$.

We sometimes integrate or differentiate power series term by term in the remainder of this proof. This is justified (on the interior of the interval of convergence) by Theorem 9.8, but we will not reference this theorem every time it is used.

(a) This follows directly from Theorem 8.6.

(b) Set $u = -x$. By part (a) we know $\frac{1}{1-u} = \sum_{n=0}^{\infty} u^n = \sum_{n=0}^{\infty}(-1)^n x^n$ when $|-x| = |x| < 1$.

(c) Since $\int \frac{1}{1+x}dx = \ln(1+x) + C$ it follows that, integrating the series for $\frac{1}{1+x}$, we have $\ln(1+x) = k + \sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}$ for some constant k. Setting $x = 0$ we see that $k = \ln(1) = 0$, so $\ln(1+x) = \sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n} = T(x)$ on $(-1,1)$.

By the Alternating Series Test, this series converges when $x = 1$ and $|T_n(x) - T(x)| < \frac{1}{n}$ on $[0,1]$. From this we conclude $\{T_n(x)\}$ converges uniformly on $[0,1]$, so by the observation above we see that $T(1) = \ln(2)$ and $T(x) = \ln(1+x)$ on $(-1,1]$.

(d) Set $u = x^2$. By part (b) we know $\frac{1}{1+u} = \sum_{n=0}^{\infty}(-1)^n u^n = \sum_{n=0}^{\infty}(-1)^n x^{2n}$ if $|x^2| < 1$

which is true on $(-1,1)$. Integrating, we have $\int \frac{1}{1+x^2}dx = k + \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{2n+1}$. Setting $x=0$ we see that $k=\tan^{-1}(0)=0$ so $\tan^{-1}(x)=\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{2n+1}$ on $(-1,1)$.

By the Alternating Series Test this series converges at both -1 and 1. Furthermore, $|s_n - T(x)| = |T_{2n-1}(x)-T(x)| \leq \frac{1}{2n+1}$ on $[-1,1]$ (where s_n is the nth partial sum of the Taylor series) so the convergence of $\{T_n(x)\}$ is uniform on $[-1,1]$ so $T(x)=\tan^{-1}(x)$ for all $x\in[-1,1]$ by the observation above.

(e) We take derivatives to get $f'(x) = \alpha(1+x)^{\alpha-1}$, $f''(x) = \alpha(\alpha-1)(1+x)^{\alpha-2}$, and so on with $f^{(n)}(x) = \alpha(\alpha-1)...(\alpha-n+1)(1+x)^{\alpha-n}$, which gives us that $T(x) = \sum_{n=0}^{\infty}\frac{\alpha(\alpha-1)(\alpha-2)...(\alpha-n+1)}{n!}x^n$, and $T(x)=f(x)$ when $T(x)$ is convergent and $\{R_n(x)\}\to 0$. Since $\lim_{n\to\infty}|\frac{\alpha(\alpha-1)...(\alpha-n)x^{n+1}}{(n+1)!}\frac{n!}{\alpha(\alpha-1)...(\alpha-n+1)x^n}|$ $=\lim_{n\to\infty}|\frac{(\alpha-n)}{(n+1)}x| = |x|$, the series $T(x)$ converges if $|x|<1$ and diverges if $|x|>1$ by the Ratio Test, but this does not show that the value $T(x)$ converges to is necessarily equal to $f(x)$.

Note that the Taylor remainder becomes zero for large values of n (since $f^{(n)}$ is zero) regardless of the choice of x if α is a positive integer.

Assume α is not a positive integer. By Taylor's Theorem (second form) we know that $|R_n(x)| \leq \frac{(\max|f^{(n+1)}(t)| \text{ on } [0,x])|x|^{n+1}}{(n+1)!}$, where $|f^{(n+1)}(t)| = |\alpha(\alpha-1)...(\alpha-n)(1+t)^{\alpha-n-1}|$. For all integers $n>\alpha$, for $x\geq 0$ the maximum of $|f^{(n+1)}(t)|$ is $|\alpha(\alpha-1)...(\alpha-n)|$. Thus, if $x\geq 0$ then we have $|R_n(x)| \leq \frac{|\alpha(\alpha-1)...(\alpha-n)||x|^{n+1}}{(n+1)!}$. We know that $\lim_{n\to\infty}\frac{x|\alpha-n|}{n+1} = x < 1$, so choose a $u\in(0,1)$ and an integer $k>\alpha$ so that if $i\geq k$ then $\frac{x|\alpha-i|}{i+1} < u$. Then for all $n>k$ we have $|R_n(x)| \leq \frac{|\alpha(\alpha-1)...(\alpha-n)||x|^{n+1}}{(n+1)!} \leq |\frac{\alpha}{1}\frac{\alpha-1}{2}...\frac{\alpha-k}{k+1}|u^{n-k}$, which approaches zero as n approaches infinity.

If $-1<x<0$ then we first observe that $\lim_{n\to\infty}n\binom{\alpha}{n}x^n = 0$. This is because

$$\lim_{n\to\infty}|\frac{(n+1)\alpha(\alpha-1)...(\alpha-n)}{(n+1)!x^n}\frac{n!}{(n)\alpha(\alpha-1)...(\alpha-n+1)}x^{n+1}| = \lim_{n\to\infty}|\frac{n+1}{n}\frac{(\alpha-n)}{(n+1)}x| =$$

$|x|<1$. Thus, by Theorem 8.12, $\lim_{n\to\infty}n\binom{\alpha}{n}x^n = 0$.

Next, we observe that $R_n(x) = \frac{1}{n!}\int_0^x f^{(n+1)}(t)(x-t)^n dt = (n+1)\binom{\alpha}{n+1}\int_0^x (1+t)^{\alpha-n-l}(x-t))^n dt$. To see this, differentiate $(1+t)^{\alpha}$ $n+1$ times and notice that $\alpha(\alpha-1)...(\alpha-n)$ can be written as $\frac{(n+1)!\alpha(\alpha-1)...(\alpha-n)}{(n+1)!}$ which means that $\frac{\alpha(\alpha-1)...(\alpha-n)}{n!} = (n+1)\binom{\alpha}{n+1}$.

Finally, we observe that $|R_n(x)| = (n+1)\binom{\alpha}{n+1}\int_x^0 (\frac{t-x}{1+t})^n(1+t)^{\alpha-1}dt$. However, if $-1 < x \leq t \leq 0$ then $\frac{t-x}{1+t} \leq |x|$ since if we differentiate this fraction with respect to t we get $\frac{(1+t)-(t-x)}{(1+t)^2} > 0$ so the maximum value occurs at $t = 0$ and is $-x$. From this we conclude that $|(\frac{t-x}{1+t})^n| \leq |x|^n$ on $[x, 0]$. Since $(1+t)^{\alpha-1}$ is continuous on $[x, 0]$, it takes on a maximum value M on that interval (which does not depend on n). Thus, $|R_n(x)| \leq M(n+1)\binom{\alpha}{n+1}\int_x^0 |x|^n dt \leq M(n+1)\binom{\alpha}{n+1}|x|^n$. By the observation above, $\lim_{n\to\infty}(n+1)\binom{\alpha}{n+1}|x|^n = 0$, so $|R_n(x)| \to 0$.

(f) We know that the nth derivative of e^x is e^x and $e^0 = 1$, so by Taylor's Theorem, it follows, for each $n \in \mathbb{N}$, that $e^x = \sum_{i=0}^{n} \frac{x^i}{i!} + \frac{1}{n!}\int_0^x e^t(x-t)^n dt$. Also, we know that since e^x is an increasing function, if $x < 0$ then $\frac{1}{n!}|\int_0^x e^t(x-t)^n dt| \leq \frac{|x|^{n+1}}{(n+1)!}$, and if $x > 0$ then $\frac{1}{n!}\int_0^x e^t(x-t)^n dt \leq \frac{e^x x^{n+1}}{(n+1)!}$. Both of these approach zero since $\lim_{n\to\infty}\frac{x^{n+1}}{(n+1)!} = 0$. Thus, $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$ on $(-\infty, \infty)$.

(g) The derivatives of $f(x) = \sin(x)$ follow the pattern $f(0) = \sin(0) = 0$, $f'(0) = \cos(0) = 1$, $f''(0) = -\sin(0) = 0$, $f'''(0) = -\cos(0) = -1$, $f^{(4)}(0) = \sin(0) = 0$ and so on. Thus, multiplying each nth derivative at zero by $\frac{x^n}{n!}$ and adding the results as normal, we end up with a Taylor series $T(x) = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ....$ For this series, since the nth derivative of $\sin(x)$ has absolute value no larger than one, it follows that $|R_n(x)| \leq \frac{(1)x^{n+1}}{(n+1)!}$ which converges to zero regardless of what x is, so $T(x) = \sin(x)$ for all real x.

(h) We differentiate each term of the Taylor series for $\sin(x)$ to get that the stated series is the series for $\cos(x)$ for all real numbers.

(i) Since the series for e^x and e^{-x} converge everywhere we can obtain the series for $\frac{e^x - e^{-x}}{2}$ by subtracting the terms of these series and dividing by two, giving us $\sinh(x) = \sum_{n=0}^{\infty}\frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + ...$ on $(-\infty, \infty)$.

(j) Differentiating the preceding series term by term gives $\cosh(x) = \sum_{n=0}^{\infty}\frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + ...$ on $(-\infty, \infty)$.

□

Note that in part (e), the series actually converges to the function $(1+x)^\alpha$ at both end points if $\alpha > 0$ and at $x = 1$ if $-1 < \alpha < 0$.

Exercises:

Exercise 9.1. *Let $\{x_n\}$ be bounded and let $\{y_n\} \to c \geq 0$. Then* $\limsup\{y_n x_n\} = c \limsup\{x_n\}$.

Exercise 9.2. *Let $\{x_n\} \to p$. Then* $\limsup\{x_n\} = p$.

Exercise 9.3. *Let $\{x_n\}$ have no upper bound and let $c > 0$. Then* $\limsup\{cx_n\} = \infty$

The next theorem actually uses the methods of sections five and six, but is useful here.

Exercise 9.4. *Prove that $\lim\limits_{n\to\infty}(n+j)^{\frac{1}{n}} = 1$ (if $n+j > 0$ for all $n \in \mathbb{N}$).*

Exercise 9.5. *Let $f(x) = e^{-\frac{1}{x^2}}$ if $x \neq 0$ and let $f(0) = 0$. Prove that f has derivatives of all orders and $f^{(n)}(0) = 0$ for all non-negative integers n, but f is not analytic on any open interval containing 0.*

Exercise 9.6. *Give an example of a sequence of functions $\{f_n\} \to f$ on a set E so that each f_n is bounded, but f is not bounded.*

Exercise 9.7. *Prove that if a sequence of functions $\{f_n\} \to f$ uniformly on a set E and each f_n is bounded then f is bounded.*

Exercise 9.8. *Give an example of a sequence of functions $\{f_n\} \to f$ on $[0,1]$ so that each f_n is bounded and continuous except on a finite set, and f is bounded, but is not continuous at any point.*

Exercise 9.9. *Let $\{f_n\} \to f$ uniformly on $[a,b]$ and let $\{g_n\} \to g$ uniformly on $[a,b]$. Then $\{f_n + g_n\} \to f + g$ uniformly on $[a,b]$. In particular, if $g_n = c_n$ and $\{c_n\} \to c$ then $\{f_n + c_n\} \to f + c$ uniformly on $[a,b]$.*

Solutions:

Solution to Exercise 9.1. *Let* $\{x_n\}$ *be bounded and let* $\{y_n\} \to c \geq 0$. *Then* $\limsup\{y_n x_n\} = c \limsup\{x_n\}$.

Proof. First, note that $\sup\{x_n, x_{n+1}, x_{n+2}, ...\}$ exists for each $n \in \mathbb{N}$ (and is always greater than or equal to $\inf\{x_1, x_2, x_3, ...\}$) since $\{x_n\}$ is bounded. Also, by an earlier exercise we know that if $n > m$ then $\sup\{x_n, x_{n+1}, x_{n+2}, ...\} \leq \sup\{x_m, x_{m+1}, x_{n+2}, ...\}$, which means that $\{\sup\{x_n, x_{n+1}, x_{n+2}, ...\}\}$ is a non-increasing sequence which is bounded below and converges. Also, we know that since $\{x_n\}$ is bounded and $\{y_n\}$ converges, that both sequences are bounded, so we can find $M, N > 0$ so that $|x_n| < M$ and $|y_n| < N$ for all $n \in \mathbb{N}$, which means that $|x_n y_n| < MN$ for all $n \in \mathbb{N}$, so $\{x_n y_n\}$ is bounded and $\limsup\{y_n x_n\}$ exists.

Set $\limsup\{x_n\} = L$, and choose $B > 0$ so that $|x_n| < B$ for all $n \in \mathbb{N}$. Let $\delta > 0$ and then choose $j \in \mathbb{N}$ so that if $n \geq j$ then $|y_n - c| < \delta$. Then $cx_n - \delta|x_n| \leq x_n y_n \leq cx_n + \delta|x_n|$. Thus, $\sup\{x_n y_n, x_{n+1}y_{n+1}, x_{n+2}y_{n+2}, ...\} \leq \sup\{cx_n + \delta B, cx_{n+1} + \delta B, ...\} \leq c\sup\{x_n, x_{n+1}, x_{n+2}, ...\} + \delta B$ by earlier theorems. Thus, $\lim_{n\to\infty} \sup\{x_n y_n, x_{n+1}y_{n+1}, x_{n+2}y_{n+2}, ...\} \leq cL$. Similarly, $\lim_{n\to\infty} \sup\{x_n y_n, x_{n+1}y_{n+1}, ...\} \geq cL$, so the result follows.

□

Solution to Exercise 9.2. *Let* $\{x_n\} \to p$. *Then* $\limsup\{x_n\} = p$.

Proof. Using exercise 9.1, this result follows immediately, setting the bounded sequence to be $\{1\}$ in the theorem statement.

A direct proof is also straightforward however:

Let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|x_n - p| < \dfrac{\epsilon}{2}$, so $p - \epsilon < \sup\{x_n, x_{n+1}, x_{n+2}, ...\} \leq p + \dfrac{\epsilon}{2} < p + \epsilon$, so $|\sup\{x_n, x_{n+1}, x_{n+2}, ...\} - p| < \epsilon$. Thus, $\{\sup\{x_n, x_{n+1}, x_{n+2}, ...\}\} \to p$. □

Solution to Exercise 9.3. *Let* $\{x_n\}$ *have no upper bound and let* $c > 0$. *Then* $\limsup\{cx_n\} = \infty$

Proof. Let $M > 0$. There is an n so that $x_n > \dfrac{M}{c}$, which means that $cx_n > M$. Thus, $\{cx_n\}$ is not bounded above, so $\limsup\{cx_n\} = \infty$.

□

Solution to Exercise 9.4. *Prove that* $\lim_{n\to\infty}(n+j)^{\frac{1}{n}} = 1$ *if* $n + j > 0$.

Proof. Taking the log we have $\lim_{n\to\infty} \ln((n+j)^(\dfrac{1}{n}) = \lim_{n\to\infty} \dfrac{1}{n}\ln(n+j)$. Using L'Hosptial's Rule, this is $\lim_{n\to\infty} \dfrac{1}{n+j} = 0$. Since e^x is continuous and the inverse of $\ln(x)$ we have that $\lim_{n\to\infty} \exp(\ln((n+j)^{\frac{1}{n}})) = \lim_{n\to\infty}(n+j)^{\frac{1}{n}} = e^0 = 1$. □

Solution to Exercise 9.5. *Let $f(x) = e^{-\frac{1}{x^2}}$ if $x \neq 0$ and let $f(0) = 0$. Prove that f has derivatives of all orders, but is not analytic on any open interval containing* 0.

Proof. Each derivative of $f^{(n)}(x)$ is a rational function multiplied by $e^{\frac{-1}{x^2}}$ everywhere except at $x = 0$. To see this, note that this is true for $n = 0$ and if the kth derivative is $\dfrac{p(x)}{q(x)}e^{\frac{-1}{x^2}}$ then $f^{(k+1)}(x) = (\dfrac{p(x)}{q(x)}\dfrac{2}{x^3} + \dfrac{q(x)p'(x) - p(x)q'(x)}{(q(x))^2})e^{-\frac{1}{x^2}}$, so the statement follows inductively.

At $x = 0$, let $f^{(n)}(x) = \dfrac{p(x)}{q(x)}e^{-\frac{1}{x^2}}$ for $x \neq 0$. Then we have $f^{(n)}(0) = \lim\limits_{x \to 0} \dfrac{\frac{p(x)}{q(x)}e^{-\frac{1}{x^2}} - 0}{x - 0}$ $= \dfrac{p(x)}{xq(x)}e^{-\frac{1}{x^2}}$, which is a rational function times $e^{-\frac{1}{x^2}}$. For any positive integer m is true that $\lim\limits_{x \to 0^+} \dfrac{1}{x^m}e^{-\frac{1}{x^2}} = \lim\limits_{u \to \infty} u^m e^{-u^2}$ by Theorem 4.14. By L'Hospital's rule we know that $\lim\limits_{u \to \infty} \dfrac{u^m}{e^u} = 0$ since differentiating the numerator m times leaves a derivative of $m!$ and differentiating the denominator still leaves e^u and $\lim\limits_{u \to \infty} \dfrac{m!}{e^u} = 0$. Since $u^m e^{-u^2} < \dfrac{u^m}{e^u}$ for all $u > 1$ it follows that $\lim\limits_{u \to \infty} u^m e^{-u^2} = 0$ by the Squeeze Theorem for extended real numbers. Similarly, it follows that $\lim\limits_{x \to 0^-} \dfrac{1}{x^m}e^{-\frac{1}{x^2}} = 0$.

If we choose m larger than the degree of $xq(x)$ then $\lim\limits_{x \to 0} \dfrac{x^m}{xq(x)} = 0$ since if k is the power of the lowest power summand Cx^k in $xp(x)$ then dividing the numerator and denominator by x^k gives us a numerator x^{m-k} which approaches zero, and a denominator which approaches C as x approaches zero. Since $0 < \dfrac{1}{xq(x)}e^{\frac{-1}{x^2}} < \dfrac{1}{x^m}e^{-\frac{1}{x^2}}$ for x sufficiently close to zero, it follows from the Squeeze Theorem that $\lim\limits_{x \to 0} \dfrac{1}{xq(x)}e^{-\frac{1}{x^2}} = 0$. We also know that $\lim\limits_{x \to 0} p(x) = p(0)$. Thus, by the product rule for limits $f^{(n)}(0) = \lim\limits_{x \to 0} \dfrac{p(x)}{xq(x)}e^{-\frac{1}{x^2}} = p(0)(0) = 0$.

Thus, all derivatives of all orders for $f(x)$ are zero at $x = 0$, so the Maclaurin series for $f(x)$ is valid only at a single point. □

Solution to Exercise 9.6. *Give an example of a sequence of functions $\{f_n\} \to f$ on a set E so that each f_n is bounded, but f is not bounded.*

Proof. Let $f_n(x) = \dfrac{1}{x + \frac{1}{n}}$ on $(0, 1]$. Each function is bounded between 0 and n, but the limit of the sequence of functions is $f(x) = \dfrac{1}{x}$ which is not bounded. □

Solution to Exercise 9.7. *Prove that if a sequence of functions $\{f_n\} \to f$ uniformly on a set E and each f_n is bounded then f is bounded.*

Proof. Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|f(x) - f_k(x)| < 1$ for all $x \in E$. Choose $M > 0$ so that $|f_k(x)| < M$ for all $x \in E$. Then $|f(x)| < M + 1$ for all $x \in E$, so f is bounded. □

Solution to Exercise 9.8. *Give an example of a sequence of functions* $\{f_n\} \to f$ *on* $[0, 1]$ *so that each* f_n *is bounded and continuous except on a finite set, and* f *is bounded, but is not continuous at any point.*

Proof. Order the set of rational numbers in $[0, 1]$ as $\mathbb{Q} = \{q_1, q_2, q_3, ...\}$. Then define $f_n(x) = 1$ is $x \in \{q_1, q_2, q_3, ..., q_n\}$ and $f_n(x) = 0$ otherwise. Then $\{f_n\} \to f$ where $f(x) = 1 if x \in \mathbb{Q}$ and $f(x) = 0$ otherwise, which is a function which is bounded and everywhere discontinuous. □

Solution to Exercise 9.9. *Let* $\{f_n\} \to f$ *uniformly on* $[a, b]$ *and let* $\{g_n\} \to g$ *uniformly on* $[a, b]$*. Then* $\{f_n + g_n\} \to f + g$ *uniformly on* $[a, b]$*. In particular, if* $g_n = c_n$ *and* $\{c_n\} \to c$ *then* $\{f_n + c_n\} \to f + c$ *uniformly on* $[a, b]$*.*

Proof. Let $\epsilon > 0$. Choose $k_1, k_2 \in \mathbb{N}$ so that if $n \geq k_1$ then $|f_n(x) - f(x)| < \dfrac{\epsilon}{2}$ for all $x \in [a, b]$ and if $n \geq k_2$ then $|g_n(x) - g(x)| < \dfrac{\epsilon}{2}$ for all $x \in [a, b]$. Then if $n \geq \max\{k_1, k_2\}$ it follows that $|(f_n(x) + g_n(x)) - (f(x) - g(x))| \leq |f_n(x) - f(x)| + |g_n(x) - g(x)| < \dfrac{\epsilon}{2} + \dfrac{\epsilon}{2} = \epsilon$, so $\{f_n + g_n\} \to f + g$ uniformly on $[a, b]$.

Finally, let $g_n(x) = c_n$ on $[a, b]$ for each $n \in \mathbb{N}$ and let $g(x) = c$ on $[a, b]$, where $\{c_n\} \to c$. Then for any $\epsilon > 0$ we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $|c_n - c| = |g_n(x) - g(x)| < \epsilon$ (for all $x \in [a, b]$), so from the previous paragraph we see that $\{f_n(x) + c_n\} \to f(x) + c$ uniformly. □

Chapter 10

Structure of Euclidean Space

We are assuming that the reader is familiar with linear algebra at this point, but many readers are not. We have placed a section on matrices in the Supplementary Materials for Multiple Variables that introduces the relatively small amount of linear algebra needed for this text.

Definition 66

The space $\mathbb{R}^n$ consists of all points which are n-tuples $(x_1, x_2, ..., x_n)$ in the n-fold Cartesian product $\mathbb{R} \times \mathbb{R} \times ... \times \mathbb{R}$. A *vector* in $\mathbb{R}^n$ is a directed line segment. We use the notation $\mathbf{x}$ to denote the vector $< x_1, x_2, ..., x_n >$ which is the directed line segment which, if placed with its base (beginning point) at the origin, its end will be the point $(x_1, x_2, x_3, ..., x_n)$. We consider all translations of $\mathbf{x}$ to be equivalent vectors to $\mathbf{x}$, all of which are written as $\mathbf{x}$. If there is no context to indicate which translation (determined by which starting point) for a vector is to be used, the vector symbol normally indicates the vector based at the origin. We interchangeably use the notation $\mathbf{x}$ to refer to the point $(x_1, x_2, x_3, ..., x_n) \in \mathbb{R}^n$, the vector $< x_1, x_2, ..., x_n >$ or the row or column matrix whose entries are the entries of the vector $< x_1, x_2, ..., x_n >$, depending on context. Furthermore, we may also use the notation $(x_1, x_2, ..., x_n)$ to refer to the vector rather than the point and $< x_1, x_2, ..., x_n >$ to refer to the point rather than the vector. We add two vectors using the convention for vectors $\mathbf{x} =< x_1, x_2, ..., x_n >$ and $\mathbf{y} =< y_1, y_2, ..., y_n >$ and real numbers (also referred to as scalars) α, β, we define $\alpha\mathbf{x} + \beta\mathbf{y} =< \alpha x_1 + \beta y_1, \alpha x_2 + \beta y_2, ..., \alpha x_n + \beta y_n >$. More formally, the n-tuple vector $\mathbf{x}$ can refer to any line segment of the form $L(\mathbf{a}, \mathbf{a}+\mathbf{x}) = \{\mathbf{a} + t\mathbf{x} \in \mathbb{R}^n | t \in [0, 1]\}$. The base of the vector is $\mathbf{a}$, obtained by plugging in $t = 0$, and the terminal point or end of the vector is $\mathbf{a} + \mathbf{x}$, obtained by plugging in $t = 1$. If $\mathbf{x} = L(\mathbf{a}, \mathbf{a} + \mathbf{x}) = L(\mathbf{b}, \mathbf{b} + \mathbf{x})$ then we refer to these line segments as translations of one another. The fact that $\mathbf{x} =< x_1, x_2, x_3, ..., x_n >$ means that when $\mathbf{a} = \mathbf{0} =< 0, 0, 0, ..., 0 >$ in this definition, the terminal point of the vector is $< x_1, x_2, ..., x_n >$. In some books we refer to the class of all such segments as the vector and a specific directed line segment as a particular realization of the vector. We will refer to different directed line segments as the same vector if they are translations of each other. We refer to the *line* through $\mathbf{a}$ and $\mathbf{a} + \mathbf{x}$ as being $l(\mathbf{a}, \mathbf{x}) = \{\mathbf{a} + t\mathbf{x} \in \mathbb{R}^n | t \in \mathbb{R}\}$.

Notation: While there is sometimes value in distinguishing between a vector list of coordinates and a list of coordinates considering an n-tuple as a point, we will normally not do so. So, $(x_1, x_2, x_3, ..., x_n)$ could refer to the vector or point (or column or row matrix) with the listed entries, as could $< x_1, x_2, ..., x_n >$ but in this text we will usually mean the vector when we write $< x_1, x_2, ..., x_n >$. In general, if a vector is listed with a bold letter label (such as $\mathbf{p}$) in an argument then it is understood that the non-bolded letter indexed with a number i (in this case p_i) refers to the ith entry of $\mathbf{p}$ unless otherwise stated.

Definition 67

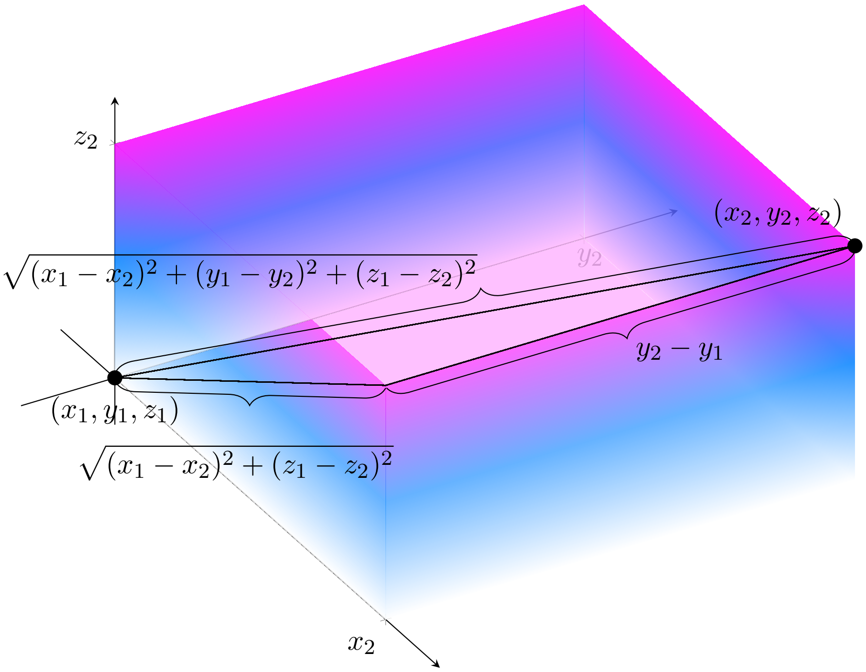
We define the norm, magnitude or length of vector $\mathbf{x} =< x_1, x_2, ..., x_n >$ to be $|\mathbf{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$. We define the *distance* from $\mathbf{x}$ to $\mathbf{y}$ to be $|\mathbf{x} - \mathbf{y}|$.

We define the *dot product* of vectors $\mathbf{x} =< x_1, x_2, ..., x_n >$ and $\mathbf{y} =< y_1, y_2, ..., y_n >$ to be $\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} a_k b_k$. Note that $\mathbf{x} \cdot \mathbf{x} = |\mathbf{x}|^2$. We define the *angle* between two vectors $\mathbf{x}$ and $\mathbf{y}$ to be $\cos^{-1}\left(\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}\right)$.

While we have defined the angle and distance as stated, there are geometric motivations for these choices, which we briefly discuss. These rely on properties of geometry that would take us a bit on a tangent, so these are more descriptions than proofs based on what we have shown. It does not do any harm to simply use these as definitions in this development, though.

The distance between two points in three dimensions can be found using the Pythagorean theorem twice, drawing a box with the points (x_1, y_1, z_1) and (x_2, y_2, z_2) at opposite corners, and for simplicity we will assume that $x_1 > x_2$. If we look at the front face of the box wherein we have a rectangle with fixed x coordinate then the diagonal of that rectangle connects (x_1, y_1, z_1) and (x_1, y_2, z_2) has length $l = \sqrt{(y_2 - y_1)^2 + (z_2 - z_1)^2}$. Then the diagonal l is perpendicular to the edge of the cube at its end point connecting (x_1, y_2, z_2) and (x_2, y_2, z_2). Thus, using the Pythagorean theorem again we get that the distance from (x_1, y_1, z_1) to (x_2, y_2, z_2) is the length of the hypotenuse of the triangle with vertices (x_1, y_1, z_1), (x_1, y_2, z_2) and (x_2, y_2, z_2), which is $d = \sqrt{(x_2 - 1_1)^2 + l^2}$. Substituting for l we motivate the following definition for distance, which can be generalized.

Distance in Three Dimensions


z_2
(x_2, y_2, z_2)
$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$
y_2
$y_2 - y_1$
(x_1, y_1, z_1)
$\sqrt{(x_1-x_2)^2+(z_1-z_2)^2}$
x_2

From this distance formula, it is possible to derive a formula for the equation for a sphere. The set of points (x, y, z) on the sphere with center (x_1, y_1, z_1) and radius r would be the set of points satisfying the equation $\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2} = r$. Squaring both sides we get the following equation for a sphere: $(x-x_1)^2+(y-y_1)^2+(z-z_1)^2 = r^2$.

Theorem 10.1. *Let* $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$*, and let* $\alpha \in \mathbb{R}$*. Then:*

(a) $\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$

(b) $\boldsymbol{u} \cdot (\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{u} \cdot \boldsymbol{w}$

(c) $(\alpha \boldsymbol{u}) \cdot \boldsymbol{v} = \alpha(\boldsymbol{u} \cdot \boldsymbol{v})$

Proof. Let $\mathbf{u} =< u_1, u_2, u_3, ..., u_n >$, $\mathbf{v} =< v_1, v_2, v_3, ..., v_n >$, and $\mathbf{w} =< w_1, w_2, w_3, ..., w_n >$.

(a) $\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3 + ... + u_nv_n = v_1u_1 + v_2u_2 + v_3u_3 + ...v_nu_n = \mathbf{v} \cdot \mathbf{u}$.

(b) $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = u_1(v_1 + w_1) + u_2(v_2 + w_2) + u_3(v_3 + w_3) + ... + u_n(v_n + w_n) = u_1v_1 + u_2v_2 + u_3v_3 + u_1w_1 + u_2w_2 + u_3w_3 + ... + u_nv_n + u_nw_n = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$.

(c) $(\alpha \mathbf{u}) \cdot \mathbf{v} = \alpha u_1v_1 + \alpha u_2v_2 + \alpha u_3v_3 + ... + \alpha u_nv_n = \alpha(u_1v_1 + u_2v_2 + u_3v_3 + ...u_nv_n)$ $= \alpha(\mathbf{u} \cdot \mathbf{v})$. □

Without the geometric development to justify things like what an angle means this argument isn't rigorous, so we will define angle in terms of dot product instead for purposes of this text.

Definition 68

Let $\mathbf{u}, \mathbf{v}$ be vectors in $\mathbb{R}^n$ we define the *angle* between $\mathbf{u}$ and $\mathbf{v}$ to be $\theta = \cos^{-1} \dfrac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$.

It would be nice to see that the angle we have defined matches what would be understood to be the angle from geometry classes. The following theorem addresses that idea. It might be a mistake to call it a theorem because we have not done the geometric development that is assumed (essentially, we are showing that our definition of angle is the same as one that has not been properly characterized with geometric axioms but which is intuitively understood by most readers).

Theorem 10.2. *Let $\boldsymbol{u}$, $\boldsymbol{v}$ be non-zero vectors (based at the origin). Then the smallest angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ is* $\cos^{-1}\left(\dfrac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\right)$.

Proof. The vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ form a triangle (with the last vector positioned to start at the end of $\mathbf{v}$ and end at the end of $\mathbf{u}$). Using the Law of Cosines, we know that if θ is the angle between $\mathbf{u}$ and $\mathbf{v}$ then $\cos(\theta) = \dfrac{|\mathbf{u}|^2 + |\mathbf{v}|^2 - |\mathbf{u} - \mathbf{v}|^2}{2|\mathbf{u}||\mathbf{v}|}$. Thus, $\cos(\theta) = \dfrac{\mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - (\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})}{2|\mathbf{u}||\mathbf{v}|} = \dfrac{\mathbf{u} \cdot \mathbf{u} + \mathbf{v} \cdot \mathbf{v} - \mathbf{u} \cdot \mathbf{u} - \mathbf{v} \cdot \mathbf{v} + 2\mathbf{u} \cdot \mathbf{v}}{2|\mathbf{u}||\mathbf{v}|}$ by Theorem 10.1, so $\cos(\theta) = \dfrac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$, as desired. □

Definition 69

Let $\mathbf{u}$, $\mathbf{v}$ be vectors in $\mathbb{R}^n$. We define the scalar projection $\text{comp}_{\mathbf{v}}\mathbf{u}$ of vector $\mathbf{u}$ onto the direction of vector $\mathbf{v}$ to be $|\mathbf{u}||\mathbf{v}|\cos(\theta)$ where θ is the smallest angle between the two vectors. Likewise, a vector in the direction of vector $\mathbf{v}$ having this length is called the vector projection of vector $\mathbf{u}$ in the direction of vector $\mathbf{v}$ and is $\mathbf{proj_v u} = \text{comp}_{\mathbf{v}}\mathbf{u}\dfrac{\mathbf{v}}{|\mathbf{v}|}$.

We say that non-zero vectors $\mathbf{u}$ and $\mathbf{v}$ are *perpendicular* or *orthogonal* if $\mathbf{u} \cdot \mathbf{v} = 0$. The zero vector is orthogonal to all vectors but is not perpendicular to any vector. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are *parallel* if $\mathbf{a} = k\mathbf{b}$ for some non-zero scalar k.

Note that two vectors are perpendicular if and only if the angle between the vectors is $\cos^{-1}(0) = \dfrac{\pi}{2}$. The plane perpendicular to a vector $< a, b, c >$ containing the point $< x_0, y_0, z_0 >$ is the set of points (x, y, z) so that $< x - x_0, y - y_0, z - z_0 >$ is perpendicular to $< a, b, c >$.

Theorem 10.3. *The plane P containing the point (x_0, y_0, z_0) which is perpendicular to the vector (a, b, c) has equation $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$.*

Proof. A point $(x, y, z) \in P$ if and only if $< x - x_0, y - y_0, z - z_0 >$ is perpendicular to $< a, b, c >$, which is true if and only if $< a, b, c > \cdot < x - x_0, y - y_0, z - z_0 >= 0$, which is true if and only if $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$. □

Definition 70

Let $\mathbf{u}$, $\mathbf{v}$ be vectors in $\mathbb{R}^n$. A *parallelogram*, two of whose sides are vectors $\mathbf{u}$ and $\mathbf{v}$, both based at one vertex $\mathbf{p}$ of the parallelogram, is the set of all points $\{\mathbf{p} + a\mathbf{u} + b\mathbf{v} + |a, b \in [0, 1]\}$. A *parallelpiped* whose edges are vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$, all based at the same vertex $\mathbf{p}$, is the set of all points $\{\mathbf{p} + a\mathbf{u} + b\mathbf{v} + c\mathbf{w}|a, b, c \in [0, 1]\}$. A *line* passing through point $\mathbf{p}$ in direction $\mathbf{u}$ (the line passing through $\mathbf{p}$ and $\mathbf{p} + \mathbf{u}$) is $l(\mathbf{p}, \mathbf{p} + \mathbf{u}) = \{\mathbf{p} + t\mathbf{u}|t \in \mathbb{R}\}$. The *angle* between lines in directions $\mathbf{u}$ and $\mathbf{v}$ is the the smallest of the angle between $\mathbf{u}$ and $\mathbf{v}$ and the angle between $\mathbf{u}$ and $-\mathbf{v}$. If the angle between two lines is zero then the lines are *parallel.* If two lines are not parallel in $\mathbb{R}^3$ and they also do not intersect then they are said to be *skew*. We use the notation $\det(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n)$ to denote the determinant of the matrix whose ith row is the vector $\mathbf{u}_i \in \mathbb{R}^n$.

If $\mathbf{u} =< u_1, u_2, u_3 >, \mathbf{v} =< v_1, v_2, v_3 >\in \mathbb{R}^3$ then the *cross product* $\mathbf{u} \times \mathbf{v} =$
$$\det \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} =< u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1 >.$$

Some of what we do in multivariable calculus will be focused specifically on $\mathbb{R}^3$ and surfaces contained in $\mathbb{R}^3$, and the cross product is helpful for many theorems in Euclidean three space. While part of the next theorem relies on geometry, we will be able to prove those results below that are dependent on geometric principles more rigorously with theorems we will prove when we cover integration.

Theorem 10.4. *Let* $\boldsymbol{u} =< u_1, u_2, u_3 >$ *and* $\boldsymbol{v} =< v_1, v_2, v_3 >$*and* $\boldsymbol{w} =< w_1, w_2, w_3 >$*. Then:*

(a) $\boldsymbol{u}\times\boldsymbol{v}$ *is the unique vector so that for every vector* $\boldsymbol{w} \in \mathbb{R}^3$*, it is true that* $\begin{vmatrix} w_1 & w_2 & w_2 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$
$= \boldsymbol{w} \cdot (\boldsymbol{u} \times \boldsymbol{v})$*.*

(b) $\boldsymbol{u} \times \boldsymbol{v} = -\boldsymbol{v} \times \boldsymbol{u}$

(c) $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are perpendicular to* $\boldsymbol{u} \times \boldsymbol{v}$

(d) $|\boldsymbol{u} \times \boldsymbol{v}| = |\boldsymbol{u}||\boldsymbol{v}| \sin(\theta)$ *where* θ *is the angle between* $\boldsymbol{u}$ *and* $\boldsymbol{v}$*.*

(e) The area of a parallelogram, two of whose sides are three coordinate vectors vectors $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *both based at one vertex of the parallelogram, is* $|\boldsymbol{u} \times \boldsymbol{v}|$*. If* $\boldsymbol{u}, \boldsymbol{v}$ *are vectors in* $\mathbb{R}^2$ *then* $|\det(\boldsymbol{u}, \boldsymbol{v})|$ *is the area of the parallelogram with those vectors as edges.*

(f) The volume of a parallelpiped, three of whose edges are vectors $\boldsymbol{u}$*,* $\boldsymbol{v}$ *and* $\boldsymbol{w}$*, all based at the same vertex of the parallelpiped, is* $|\boldsymbol{w} \cdot (\boldsymbol{u} \times \boldsymbol{v})| = |\det(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v})|$*.*

(g) If $\boldsymbol{w} = \boldsymbol{u} \times \boldsymbol{v}$ *then* $\boldsymbol{v} = \boldsymbol{w} \times \boldsymbol{u}$*.*

Proof. (a) Simply using the definition, $\begin{vmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_1 & v_3 \end{vmatrix} = w_1 \begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} - w_2 \begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} + w_3 \begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix}$ $= w_1(u_2v_3 - u_3v_2) + w_2(u_3v_1 - u_1v_3) + w_3(u_1v_2 - u_2v_1) = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$. The fact that this is the only vector having this property follows from taking the dot products of $\mathbf{u} \times \mathbf{v}$ with the component vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ since these dot products would give us that the first, second and third components of a vector having the property that for every vector $\mathbf{w} =< w_1, w_2, w_3 >\in \mathbb{R}^3$, it is the true that $\begin{vmatrix} w_1 & w_2 & w_2 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$ are exactly $u_2v_3 - u_3v_2$, $u_3v_1 - u_1v_3$ and $u_1v_2 - u_2v_1$ respectively.

(b) Since determinants are negated by switching two rows of a matrix, this follows from part (a). Another argument would be simply applying the definition, since $\mathbf{v} \times \mathbf{u} =< v_2u_3 - v_3u_2, v_3u_1 - v_1u_3, v_1u_2 - v_2u_1 > = - < u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1 >= -\mathbf{u} \times \mathbf{v}$.

(c) This follows from the fact that $\mathbf{u} \cdot \mathbf{u} \times \mathbf{v}$ is the determinant of a matrix whose first two rows are $\mathbf{u}$, which is zero. Alternately, we plug into the usual formula to get that $\mathbf{u} \cdot \mathbf{u} \times \mathbf{v} = u_1(u_2v_3 - u_3v_2) + u_2(u_3v_1 - u_1v_3) + u_3(u_1v_2 - u_2v_1) = 0$ and $\mathbf{v} \cdot \mathbf{u} \times \mathbf{v} = v_1(u_2v_3 - u_3v_2) + v_2(u_3v_1 - u_1v_3) + v_3(u_1v_2 - u_2v_1) = 0$ directly through cancellation.

(d) By definition, we have $|\mathbf{u} \times \mathbf{v}|^2 = (u_2v_3 - u_3v_2)^2 + (u_3v_1 - u_1v_3)^2 + (u_1v_2 - u_2v_1)^2$ $= u_2^2v_3^2 + u_3^2v_2^2 + u_3^2v_1^2 + u_1^2v_3^2 + u_1^2v_2^2 + u_2^2v_1^2 - 2(u_2u_3v_2v_3 + u_1u_3v_1v_3 + u_1u_2v_1v_2)$. Also, $|\mathbf{u}|^2|\mathbf{v}|^2 = (u_1^2 + u_2^2 + u_3^2)(v_1^2 + v_2^2 + v_3^2) = u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2 + u_1^2v_2^2 + u_1^2v_3^2 + u_2^2v_1^2 + u_2^2v_3^3 + u_3^2v_1^2 + u_3^2v_2^2$, and $(\mathbf{u} \cdot \mathbf{v})^2 = (u_1v_1 + u_2v_2 + u_3v_3)^2 = u_1^2v_1^2 + u_2^2v_2^2 + u_3^2v_3^2 + 2(u_2u_3v_2v_3 + u_1u_3v_1v_3 + u_1u_2v_1v_2)$. Thus, $|\mathbf{u} \times \mathbf{v}|^2 = |\mathbf{u}|^2|\mathbf{v}|^2 - (\mathbf{u} \cdot \mathbf{v})^2 = |\mathbf{u}|^2|\mathbf{v}|^2 - |\mathbf{u}|^2|\mathbf{v}|^2\cos^2(\theta) = |\mathbf{u}|^2|\mathbf{v}|^2(1 - \cos^2(\theta)) = |\mathbf{u}|^2|\mathbf{v}|^2\sin^2(\theta)$. Taking the square root of both sides gives $|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}|\sin(\theta)$.

(e) The area of a parallelogram is the product of the length of adjacent sides of the parallelogram times the sine of the angle between them. In this case, that is $|\mathbf{u}||\mathbf{v}|\sin(\theta) = |\mathbf{u} \times \mathbf{v}|$ by part (d). In the case of two coordinate vectors $< u_1, u_2 >$ and $< v_1, v_2 >$ the area of the parallelogram with those vectors as sides is the area of the parallelogram with sides $< u_1, u_2, 0 >$ and $< v_1, v_2, 0 >$, the cross product of which is $< 0, 0, u_1v_2 - u_2v_1 >$, the norm of which is $|u_1v_2 - u_2v_1| = \det(\mathbf{u}, \mathbf{v})$.

(f) Intuitively, the volume of a parallelpiped is just its height times the area of its base since every cross section parallel to the base parallelogram is a congruent parallelogram with the same area. In other words, if we view two of the vectors as vectors within the base of the parallelogram (say $\mathbf{u}$ and $\mathbf{v}$) it is the projection of the direction of the third vector $\mathbf{w}$ onto the direction perpendicular to this base ($\mathbf{u} \times \mathbf{v}$) times the area of the base parallelogram which is the area of the parallelogram. Since the area of the base is $|\mathbf{u} \times \mathbf{v}|$, we use the projection formula to get that the volume is $|\mathbf{w} \cdot \frac{\mathbf{u} \times \mathbf{v}}{|\mathbf{u} \times \mathbf{v}|}||\mathbf{u} \times \mathbf{v}| = |\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})|$ $= | \begin{vmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_1 & v_3 \end{vmatrix} | = |\det(\mathbf{w}, \mathbf{u}, \mathbf{v})|$. This result is also proven later (as a consequence Theorem 12.61, for example) without relying on the notions Cavalieri's principle addressed above.

(g) By part (a) we know that $\mathbf{v} \cdot \mathbf{w} \times \mathbf{u} = \det(\mathbf{v}, \mathbf{w}, \mathbf{u})$ and that $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = \det(\mathbf{w}, \mathbf{u}, \mathbf{v})$. Since switching rows negates the determinant, two row switches performed consecutively leaves the determinant unchanged, from which we see that $\mathbf{v} \cdot \mathbf{w} \times \mathbf{u} = \mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$.

Since some students may not have reviewed linear algebra recently, we can do additional algebra to verify that $\det(\mathbf{v},\mathbf{w},\mathbf{u}) = v_1(w_2u_3 - w_3u_2) + v_2(w_3u_1 - w_1u_3) + v_3(w_1u_2 - w_2u_1)$ $= u_1(v_2w_3 - v_3w_2) + u_2(v_3w_1 - v_1w_3) + u_3(v_1w_2 - w_1v_2) = \det(\mathbf{u},\mathbf{v},\mathbf{w})$. □

Theorem 10.5. *Cauchy-Schwarz Inequality. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then $|\boldsymbol{x} \cdot \boldsymbol{y}| \leq |\boldsymbol{x}||\boldsymbol{y}|$.*

Proof. Since $0 \leq |\mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{y}|^2}\mathbf{y}|^2 = (\mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{y}|^2}\mathbf{y}) \cdot (\mathbf{x} - \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{y}|^2}\mathbf{y}) = |\mathbf{x}|^2 - 2\frac{|\mathbf{x}\cdot\mathbf{y}|^2}{|\mathbf{y}|^2} + \frac{|\mathbf{x}\cdot\mathbf{y}|^2}{|\mathbf{y}|^4}|\mathbf{y}|^2$. Thus, $|\mathbf{x}\cdot\mathbf{y}|^2 \leq |\mathbf{x}|^2|\mathbf{y}|^2$ and $|\mathbf{x}\cdot\mathbf{y}| \leq |\mathbf{x}||\mathbf{y}|$. □

Theorem 10.6. *Triangle Inequality. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then:*

(a) $|\boldsymbol{x} + \boldsymbol{y}| \leq |\boldsymbol{x}| + |\boldsymbol{y}|$.

(b) $|\boldsymbol{x} - \boldsymbol{y}| \geq |\boldsymbol{x}| - |\boldsymbol{y}|$

(c) If $\boldsymbol{x}_i \in \mathbb{R}^n$ for all $1 \leq i \leq m$ then $\sum_{i=1}^{m} |\boldsymbol{x}_i| \geq |\sum_{i=1}^{m} \boldsymbol{x}_i|$.

Proof. (a) We know that $0 \leq |\mathbf{x}+\mathbf{y}|^2 = (\mathbf{x}+\mathbf{y})\cdot(\mathbf{x}+\mathbf{y}) = |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2\mathbf{x}\cdot\mathbf{y} \leq |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2|\mathbf{x}||\mathbf{y}| = (|\mathbf{x}|+|\mathbf{y}|)^2$ by the Cauchy-Schwarz Inequality. Thus, $|\mathbf{x}+\mathbf{y}| \leq |\mathbf{x}|+|\mathbf{y}|$.

(b) By (a) we know that $|\mathbf{y} + (\mathbf{x}-\mathbf{y})| \leq |\mathbf{y}| + |\mathbf{x}-\mathbf{y}|$ so $|\mathbf{x}-\mathbf{y}| \geq |\mathbf{x}| - |\mathbf{y}|$.

(c) We proceed by induction. This is true if $m = 1$. Assume it is true for $m = k \in \mathbb{N}$. Then $|\sum_{i=1}^{k+1} \mathbf{x}_i| = |\sum_{i=1}^{k} \mathbf{x}_i + \mathbf{x}_{k+1}| \leq |\sum_{i=1}^{k} \mathbf{x}_i| + |\mathbf{x}_{k+1}|$ by (a), which is less than or equal to $\sum_{i=1}^{k} |\mathbf{x}_i| + |\mathbf{x}_{k+1}|$ by the induction hypothesis, so $|\sum_{i=1}^{k+1} \mathbf{x}_i| \leq \sum_{i=1}^{k+1} |\mathbf{x}_i|$. The result follows for all natural numbers m. □

Some theorems are more helpful to prove in the general setting of metric spaces so that they can be used for other purposes, so we also define metric spaces before we start proving theorems about calculus in higher dimensions.

Definition 71

A set of points X with a distance function $\rho : X \times X \to [0,\infty)$ (called a *metric* for X) is a *metric space* provided that:
(1) $\rho(x,y) = 0$ if and only if $x = y$
(2) Symmetry: $\rho(x,y) = \rho(y,x)$ for all $x, y \in X$
(3) Triangle inequality: $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ for all $x, y, z \in X$.

Theorem 10.7. $\mathbb{R}^n$ *is a metric space under the metric $\rho(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|$.*

Proof. The distance norm $\rho(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|$ satisfies all of the properties of a metric. Properties (1) and (2) are immediate from the definition since the difference of two entries squared is the same no matter which term is subtracted from which, and is only zero if the entries are equal. The triangle inequality was proven, so the result follows. □

Definition 72

In a metric space (X,d) we define an *open ball* about p of radius $\epsilon > 0$ to be $B_\epsilon(p) = \{x \in X|\rho(x,p) < \epsilon\}$. We also define the closed ball about p of radius $\epsilon > 0$ to be $\overline{B_\epsilon(p)} = \{x \in X|\rho(x,p) \leq \epsilon\}$. A set A is *open* in X (or just open) if for every point $p \in A$ there is an open ball containing p which is contained in A. The complement of an open set is *closed*. We say that q is a *limit point* of set A if every open ball about q contains a point of A distinct from q. A point of A that is not a limit point of A is an *isolated point* of A. A set U is *open in* a set E if $U = V \cap E$ for some open set V, and A is *closed in* E if $A = C \cap E$ for some closed set C. A set S is *bounded* if there is some $M > 0$ so that $S \subseteq B_M(p)$ for some $p \in X$. The *diameter* of a set $S \subseteq X$ is $\sup\limits_{x,y\in S} \rho(x,y)$ if S is bounded, and infinity otherwise.

Let (X,d), (Y,ρ) be a metric spaces, $D \subset X$. A *sequence* of points in X is a function $g : \mathbb{N} \to X$, where we usually denote the image of the nth integer as a letter subscripted with n, so we might write $g(n) = x_n$ for instance. We typically use $\{x_n\}$ or $\{x_n\}_{n\in\mathbb{N}}$ as the notation for such a function g rather than the symbol g itself. We say that $\{x_n\} \to p$ in X if for every $\epsilon > 0$ there is some $k \in \mathbb{N}$ so that if $n \geq k$ then $d(x_n,p) < \epsilon$. We say that $f : D \to Y$ is *continuous* at $c \in X$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $x \in D$ and $d(x,c) < \delta$ then $\rho(f(x),f(c)) < \epsilon$. Let p be a limit point of D. We say that $\lim\limits_{x\to p} f(x) = L$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if $0 < d(x,p) < \delta$ and $x \in D$ then $\rho(f(x),L) < \epsilon$.

Note that since an open set and its complement are open and closed respectively, their intersections with E are open in E and closed in E respectively, meaning that the complement (in E) of a set which is open in E is closed in E and the complement in E of a closed set in E is open in E.

Theorem 10.8. *Let (X,d) be a metric space and let $p \in X$ and let $\epsilon > 0$. Then $B_\epsilon(p)$ is open and $\overline{B_\epsilon(p)}$ is closed.*

Proof. Let $q \in B_\epsilon(p)$. Set $\delta = \epsilon - d(p,q)$. Let $z \in B_\delta(q)$. Then $d(z,p) \leq d(z,q) + d(q,p) < \epsilon$ by the triangle inequality, so $z \in B_\epsilon(p)$, which means that $B_\delta(q) \subseteq B_\epsilon(p)$, so $B_\epsilon(p)$ is open.

Let $q \in X \setminus \overline{B_\epsilon(p)}$. Then $d(p,q) > \epsilon$. Set $\delta = d(p,q) - \epsilon$. Let $z \in B_\delta(q)$. Then $d(p,z) \geq d(p,q) - d(z,q) > \epsilon$ by the triangle inequality, so $B_\delta(q) \subseteq X \setminus \overline{B_\epsilon(p)}$, which means that $X \setminus \overline{B_\epsilon(p)}$ is open so $\overline{B_\epsilon(p)}$ is closed.

□

Theorem 10.9. *Let $\{\boldsymbol{x}_i\}$ be a sequence in $\mathbb{R}^m$, where each $\boldsymbol{x}_i = (x_{i_1}, x_{i_2}, ..., x_{i_m})$. Then $\{\boldsymbol{x}_i\} \to \boldsymbol{p} = (p_1, p_2, ..., p_m)$ if and only if $\{x_{i_j}\} \to p_j$ for each $1 \leq j \leq m$.*

Proof. First, assume that $\{\mathbf{x}_i\} \to \mathbf{p}$. Then let $\epsilon > 0$. Choose k so that if $i \geq k$ then $|\mathbf{x}_i - \mathbf{p}| < \epsilon$. Since $|\mathbf{x}_i - \mathbf{p}| \geq |x_{i_j} - p_j|$ we know that $|x_{i_j} - p_j| < \epsilon$ so $\{x_{i_j}\} \to p_j$.

Conversely, if we assume $\{x_{i_j}\} \to p_j$ for each integer $j \in \{0, 1, 2, ..., m\}$. Then we can choose $k \in \mathbb{N}$ so that if $i \geq k$ then $|x_{i_j} - p_j| < \dfrac{\epsilon}{\sqrt{m}}$ for each $1 \leq j \leq m$. Thus, $|\mathbf{x}_i - \mathbf{p}| \leq \sqrt{\displaystyle\sum_{i=1}^{m} \frac{\epsilon^2}{m}} = \epsilon$. Hence, $\{\mathbf{x}_i\} \to \mathbf{p}$. □

Theorem 10.10. *Let (X, d) be a metric space with $c \in X$, and let $f : D \to Y$ be a function, where (Y, ρ) is a metric space and $D \subseteq X$. Then $\lim\limits_{x \to c} f(x) = L$ if and only if $\lim\limits_{x \to c} \rho(f(x), L) = 0$.*

Proof. Let $\epsilon > 0$. First, assume that $\lim\limits_{x\to c} \rho(f(x), L) = 0$. Since the function $\rho(f(x), L) : D \to \mathbb{R}$ has a limit at c, it follows that c is a limit point of D. Choose $\delta > 0$ so that if $0 < d(x, c) < \delta$ then $|\rho(f(x), L) - 0| < \epsilon$. Then it follows that $\rho(f(x), L) < \epsilon$, so $\lim\limits_{x\to c} f(x) = L$.

Assume $\lim\limits_{x\to c} f(x) = L$. Then c is a limit point of X. Choose $\delta > 0$ so that if $0 < d(x, c) < \delta$ then $\rho(f(x), L) < \epsilon$. Then $|\rho(f(x), L) - 0| < \epsilon$, so $\lim\limits_{x\to c} \rho(f(x), L) = 0$. □

Theorem 10.11. *Squeeze Theorem (for metric spaces)*

Let $f, g : D \to Y$ where (X, d) and (Y, ρ) are metric spaces and $D \subseteq X$.

(a) If $\lim\limits_{x\to c} \rho(g(x), L) = 0$ and $\rho(f(x), L) \leq \rho(g(x), L)$ for all $x \in B_\delta(c)$ for some $\delta > 0$ then $\lim\limits_{x\to c} f(x) = L$.

(b) If $\{x_n\} \to 0$ in $\mathbb{R}$ and $d(z_n, p) \leq x_n$ for all $n \geq k$ for some $k \in \mathbb{N}$ then $\{z_n\} \to p$ in X.

Proof. (a) We know c is a limit point of D since $\lim\limits_{x\to c} g(x) = L$. Let $\epsilon > 0$. We can choose a $0 < \gamma < \delta$ so that if $d(x, c) < \gamma$ then $\rho(g(x), L) < \epsilon$ for all $x \in D$, which implies that $\rho(f(x), L) < \epsilon$ so $\lim\limits_{x\to c} f(x) = L$.

(b) Let $\epsilon > 0$. Choose $t \in \mathbb{N}$ so that $t > k$ and if $n \geq t$ then $|x_n - 0| < \epsilon$. Then if $n \geq t$ we know that $d(z_n, p) < x_n < \epsilon$ so $\{z_n\} \to p$. □

Theorem 10.12. *Let A be a subset of a metric space X. Then p is a limit point of A if and only if there is a sequence of points in $A \setminus \{p\}$ converging to p.*

Proof. First, assume there is a such a sequence. Then for every $\epsilon > 0$, for sufficiently large n we know that $x_n \in B_\epsilon(p)$, which means that p is a limit point of A.

Next, assume that p is a limit point of A. Then for each n we can choose a point $p_n \in B_{\frac{1}{n}}(p) \cap A \setminus \{p\}$. Then $\{p_n\} \to p$ by the Squeeze Theorem. □

Theorem 10.13. *The Sequential Characterization of Limits for functions on metric spaces. Let $f : D \to Y$, a metric space with metric ρ, where $D \subseteq X$, a metric space with metric d, and c is a limit point of D. Then $\lim\limits_{x \to c} f(x) = L$ if and only if for every sequence $\{x_n\} \subseteq D \setminus \{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$.*

Proof. First, assume that $\lim\limits_{x \to c} f(x) = L$. Let $\{x_n\} \subseteq D \setminus \{c\}$ such that $\{x_n\} \to c$. Let $\epsilon > 0$. Then for some $\delta > 0$, we know that if $0 < d(x,c) < \delta$ and $x \in D$ then $\rho(f(x), L) < \epsilon$. Since $\{x_n\} \to c$, we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $d(x_n, c) < \delta$, and since $\{x_n\} \subseteq D \setminus \{c\}$ it follows that if $n \geq k$ then $0 < d(x_n, c) < \delta$. Hence, if $n \geq k$ then $\rho(f(x_n), L) < \epsilon$, so $\{f(x_n)\} \to L$.

Next, assume that for every sequence $\{x_n\} \subseteq D \setminus \{c\}$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to L$. Suppose that it is false that $\lim\limits_{x \to c} f(x) = L$. Then we can find an $\epsilon > 0$ so that for every $\delta > 0$ there is some $x \in D \setminus \{c\}$ so that $d(x, c) < \delta$ but $\rho(f(x), L) \geq \epsilon$. For each $n \in \mathbb{N}$ we choose $x_n \in D \setminus \{c\}$ so that $d(x_n, c) < \dfrac{1}{n}$ and $\rho(f(x_n), L) \geq \epsilon$. Then $\{x_n\} \to c$ by the Squeeze Theorem, but $\{f(x_n)\} \not\to L$, contradicting our assumption.

□

Theorem 10.14. *The Sequential Characterization of Continuity for functions on metric spaces. Let $f : D \to Y$, a metric space with metric ρ, where $D \subseteq X$, a metric space with metric d, and $c \in D$. Then f is continuous at c if and only if for every sequence $\{x_n\} \subseteq D$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to f(c)$.*

Proof. First, assume that f is continuous at c. Let $\{x_n\} \subseteq D$ such that $\{x_n\} \to c$. Let $\epsilon > 0$. Then for some $\delta > 0$, we know that if $d(x, c) < \delta$ and $x \in D$ then $\rho(f(x), f(c)) < \epsilon$. Since $\{x_n\} \to c$, we can find $k \in \mathbb{N}$ so that if $n \geq k$ then $d(x_n, c) < \delta$. Hence, if $n \geq k$ then $\rho(f(x_n), f(c)) < \epsilon$, so $\{f(x_n)\} \to f(c)$.

Next, assume that for every sequence $\{x_n\} \subseteq D$, if $\{x_n\} \to c$ then $\{f(x_n)\} \to f(c)$. Suppose that f is not continuous at c. Then we can find an $\epsilon > 0$ so that for every $\delta > 0$ there is some $x \in D$ so that $d(x, c) < \delta$ but $\rho(f(x), f(c)) \geq \epsilon$. For each $n \in \mathbb{N}$ we choose $x_n \in D$ so that $d(x_n, c) < \dfrac{1}{n}$ and $\rho(f(x_n), f(c)) \geq \epsilon$. Then $\{x_n\} \to c$ by the Squeeze Theorem, but $\{f(x_n)\} \not\to f(c)$, contradicting our assumption.

□

Theorem 10.15. *Let $f : D \to Y$, where $D \subseteq X$, a metric space with metric d, Y is a metric space with distance function ρ and $c \in D$.*

(a) Let c be a limit point of D. Then f is continuous at c if and only if $\lim\limits_{x \to c} f(x) = f(c)$.

(b) If c is an isolated point of D then f is continuous at c.

Proof. (a) First, assume that f is continuous at c. Let $\epsilon > 0$. We know that for some $\delta > 0$, if $d(x, c) < \delta$ and $x \in D$ then $\rho(f(x), f(c)) < \epsilon$. Hence, if $0 < d(x, c) < \delta$ and $x \in D$ then $\rho(f(x), f(c)) < \epsilon$, so $\lim\limits_{x \to c} f(x) = f(c)$.

Next, assume that $\lim_{x\to c} f(x) = f(c)$. Let $\epsilon > 0$. We know that for some $\delta > 0$, if $0 < d(x, c) < \delta$ and $x \in D$ then $\rho(f(x), f(c)) < \epsilon$, but if $c = x$ then $\rho(f(x), f(c)) = 0 < \epsilon$ as well. Hence, if $d(x, c) < \delta$ and $x \in D$ then $\rho(f(x), f(c)) < \epsilon$, so f is continuous at c.

(b) Since c is an isolated point of D, we can find $\delta > 0$ so that the only point D whose distance from c is less than δ is c. Hence, if $d(x, c) < \delta$ and $x \in D$ then $x = c$, so $\rho(f(x), f(c)) = 0$ which is less than any positive number ϵ and therefore f is continuous at c.

□

Theorem 10.16. *Let $f : X \to Y$ and $g : Y \to Z$ be functions where X, Y, Z are metric and $p \in dom(g \circ f)$. Let f be continuous at p and g be continuous at $f(p)$. Then $g \circ f$ is continuous at p.*

Proof. Let $\{x_n\} \to p$, where $\{x_n\} \subseteq dom(g \circ f)$. Then $\{f(x_n)\} \to f(p)$ and so $\{g(f(x_n))\} \to g(f(p))$, which implies that $g \circ f$ is continuous at p.

□

Theorem 10.17. *Let $f : D \to \mathbb{R}^m$ where $D \subseteq \mathbb{R}^n$, and $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_m(\boldsymbol{x}))$ where each $f_i(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{c}$ is a limit point of D. Then*

(a) $\lim_{\boldsymbol{x}\to\boldsymbol{c}} f(\boldsymbol{x}) = \boldsymbol{L} = (L_1, L_2, ..., L_m)$ if and only if $\lim_{\boldsymbol{x}\to\boldsymbol{c}} f_i(\boldsymbol{x}) = L_i$ for each $1 \le i \le m$.

(b) f is continuous at $\mathbf{p} \in D$ if and only if each component function f_i is continuous at $\mathbf{p}$.

Proof. (a) By the Sequential Characterization of Limits we know that $\lim_{\mathbf{x}\to\mathbf{c}} f(\mathbf{x}) = \mathbf{L}$ if and only if for every sequence $\{\mathbf{x}_i\} \subseteq D \setminus \{c\}$ so that $\{\mathbf{x}_i\} \to \mathbf{c}$ it is true that $\{f(\mathbf{x}_i)\} \to \mathbf{L}$. Choose a sequence $\{\mathbf{x}_i\} \subseteq D \setminus \{c\}$ so that $\{\mathbf{x}_i\} \to \mathbf{c}$.

By Theorem 10.9, $\{f(\mathbf{x}_i)\} \to \mathbf{L}$ if and only if $\{f_j(\mathbf{x}_i)\} \to L_j$ for all $1 \le j \le m$.

By the Sequential Characterization of Limits, for all $1 \le j \le m$, $\{f_j(\mathbf{x}_i)\} \to L_j$ for each sequence $\{\mathbf{x}_i\} \subseteq D \setminus \{c\}$ so that $\{\mathbf{x}_i\} \to \mathbf{c}$ if and only if $\lim_{\mathbf{x}\to\mathbf{c}} f_j(\mathbf{x}) = L_j$. Hence, the result follows.

(b) By the Sequential Characterization of Continuity we know that f is continuous at $\mathbf{p}$ if and only if for every sequence $\{\mathbf{x}_i\} \subseteq D$ so that $\{\mathbf{x}_i\} \to \mathbf{p}$ it is true that $\{f(\mathbf{x}_i)\} \to f(\mathbf{p})$.

By Theorem 10.9, $\{f(\mathbf{x}_i)\} \to f(\mathbf{p})$ if and only if $\{f_j(\mathbf{x}_i)\} \to f(\mathbf{p})_j$ for all $1 \le j \le m$, where $f(\mathbf{p})_j$ denotes the jth component of $f(\mathbf{p})$.

By the Sequential Characterization of Continuity, we also know that $\{f_j(\mathbf{x}_i)\} \to f(\mathbf{p})_j$ for all $1 \le j \le m$ for every sequence $\{\mathbf{x}_i\} \subseteq D$ so that $\{\mathbf{x}_i\} \to \mathbf{p}$ if and only if f_j is continuous at $\mathbf{p}$ for all $1 \le j \le m$. Hence, the result follows.

□

Theorem 10.18. *Let (X, d) be a metric space, let $D \subseteq X$ and let c be a limit point of D. Let $g : D \to \mathbb{R}$, and $\lim_{x\to c} g(x) = t \neq 0$. Then there is a $\delta > 0$ so that if $x \in B_\delta(c) \cap D$ then $g(x) > \dfrac{|t|}{2}$, and c is a limit point of $dom(g)$.*

Proof. We can find $\delta > 0$ so that if $d(x,c) < \delta$ then $|g(x) - t| < \frac{|t|}{2}$, which means that $|g(x)| > \frac{|t|}{2}$, and therefore $\frac{1}{|g(x)|} \geq \frac{2}{|t|} > 0$. Hence if U is any open set containing c then $B_\delta(c) \cap U$ contains a point q distinct from c which is in D, and therefore in the domain of $\frac{1}{g}$ since $g(q) \neq 0$. Thus, c is a limit point of $dom(g)$. □

Theorem 10.19. *Let (X,d) be a metric space, let $D \subseteq X$ and let c be a limit point of D. Let $f,g : D \to \mathbb{R}$, and $\lim_{x\to c} g(x) = s$ and $\lim_{x\to c} g(x) = t$. Then:*

(a) $\lim_{x\to c} \alpha f(x) + \beta g(x) = \alpha s + \beta t$ for any real α, β.

(b) $\lim_{x\to c} f(x)g(x) = st$.

(c) $\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{s}{t}$ if $t \neq 0$.

Proof. Let $\{x_n\}$ be a sequence in $D\backslash\{c\}$ which converges to c. By the Sequential Characterization of Limits we know that $\{f(x_n)\} \to s$ and $\{g(x_n)\} \to t$. Hence, $\{\alpha f(x_n)+\beta g(x_n)\} \to \alpha s+\beta t$, $\{f(x_n)g(x_n)\} \to st$, and if $\{x_n\} \subseteq dom(\frac{f}{g})$ then $\{\frac{f(x_n)}{g(x_n)}\} \to \frac{s}{t}$. By Theorem 10.18, we know that c is a limit point of the domain of $\frac{f}{g}$ if $t \neq 0$, and therefore by the Sequential Characterization of Limits we conclude that (a), (b) and (c) are true.

□

Theorem 10.20. *Let (X,d) be a metric space, let $D \subseteq X$ and let $c \in D$. Let $f,g : D \to \mathbb{R}$ be continuous at c. Then fg and $f+g$ are continuous at c, and $\frac{f}{g}$ is continuous at c if $g(c) \neq 0$.*

Proof. By Theorem 10.19, we know that if c is a limit point of D then fg and $f+g$ are continuous at c by Theorem 10.15, and if c is a limit point of D and $g(c) \neq 0$ then c is a limit point of the domain of $\frac{f}{g}$ by Theorem 10.18 and so $\frac{f}{g}$ is also continuous at c.

If c is not a limit point of D, then c is not a limit point of the domains of $f+g$ or fg or $\frac{f}{g}$, which means that f, g, $f+g$, and fg are all continuous at c, and $\frac{f}{g}$ is continuous at c if $\frac{f}{g}$ is defined at c, which is true if $g(c) \neq 0$, again by Theorem 10.15. □

Theorem 10.21. *If U_α is open in a metric space X for all $\alpha \in J$ (where J is an arbitrary indexing set) then $\bigcup_{\alpha\in J} U_\alpha$ is open.*

Proof. Let $p \in \bigcup_{\alpha\in J} U_\alpha$. Then $p \in U_\beta$ for some $\beta \in J$, so for some $\epsilon > 0$ it follows that $B_\epsilon(p) \subseteq U_\beta \subseteq \bigcup_{\alpha\in J} U_\alpha$, so $\bigcup_{\alpha\in J} U_\alpha$ is open.

□

Theorem 10.22. *If $U_1, U_2, ..., U_n$ are open sets in a metric space X then $\bigcap_{i=1}^{n} U_i$ is open.*

Proof. If $p \in \bigcap_{i=1}^{n} U_i$ then for each $i \leq n$ we can find $\epsilon_i > 0$ so that $B_{\epsilon_i}(p) \subseteq U_i$. Hence, if we set $\epsilon = \min\{\epsilon_1, \epsilon_2, ..., \epsilon_n\}$ then $B_\epsilon(p) \subseteq \bigcap_{i=1}^{n} U_i$.

□

Theorem 10.23. *If A_α is closed in a metric space X for all $\alpha \in J$ (where J is an arbitrary indexing set) then $\bigcap_{\alpha \in J} A_\alpha$ is closed.*

Proof. By DeMorgan's Laws, $X \setminus \bigcap_{\alpha \in J} A_\alpha = \bigcup_{\alpha \in J} X \setminus A_\alpha$, which is open. Hence, $\bigcap_{\alpha \in J} A_\alpha$ is closed.

□

Theorem 10.24. *If $A_1, A_2, ..., A_n$ are closed subsets of a metric space X sets then $\bigcup_{i=1}^{n} A_i$ is closed.*

Proof. By DeMorgan's Laws, $X \setminus \bigcup_{i=1}^{n} A_i = \bigcap_{i=1}^{n} X \setminus A_i$, which is open. Hence, $\bigcup_{i=1}^{n} A_i$ is closed.

□

Theorem 10.25. *Let (X, ρ) be a metric space and let $A \subseteq E \subseteq X$. Then: (a) A is closed in E if and only if A contains all of its limit points which are in E.*

(b) A is open in E if and only if for every $p \in A$ there is an $\epsilon > 0$ so that $B_\epsilon(p) \cap E \subseteq A$.

(c) A is open in E if and only if A is a union of open balls in E (open balls intersected with E).

Proof. (a) First assume A is closed in E and pick a closed set C so that $C \cap E = A$. Let $p \in E \setminus A$. Then $p \notin C$. Since $X \setminus C$ is open there is an open ball $B_\epsilon(p) \subseteq X \setminus C$. Since $B_\epsilon(p) \cap C = \emptyset$, $B_\epsilon(p) \cap A = \emptyset$, so p is not a limit point of A.

Next assume that A contains all of its limit points which are contained in E and let $p \in (X \setminus A) \cap E$. Then p is not a limit point of A, which means that there is some $B_{\epsilon_p}(p)$ which contains no points of A. Choosing such a ball for every $p \in (X \setminus A) \cap E$ we obtain an open set $W = \bigcup_{p \in (X \setminus A) \cap E} B_{\epsilon_p}(p)$ which does not intersect A and contains all points of $E \setminus A$. Thus, $X \setminus W$ is closed so $E \cap (X \setminus W) = A$ is closed in E.

(b) Assume A is open in E. Then there is an open set U so that $U \cap E = A$. Let $p \in A$. Then there is an $\epsilon > 0$ so that $B_\epsilon(p) \subseteq U$, which means that $B_\epsilon(p) \cap E \subseteq A$.

Assume that for every $p \in A$ there is an $\epsilon > 0$ so that $B_{\epsilon_p}(p) \cap E \subseteq A$. Let $W = \bigcup_{p \in A} B_{\epsilon_p}(p)$. Then W is open and $W \cap E = A$, which means that A is open in E.

(c) First, let A be open in E. For each $x \in A$ by part (b) we can choose $\epsilon_x > 0$ so that $B_{\epsilon_x}(x) \cap E \subseteq A$. Then since each point of A is the center of such an open ball, we know

$\bigcup_{x \in A} B_{\epsilon_x}(x) \cap E \supseteq A$. However, since each $B_{\epsilon_x}(x) \cap E$ is a subset of A we also know that $\bigcup_{x \in A} (B_{\epsilon_x}(x) \cap E) \subseteq A$. Hence, $\bigcup_{x \in A} (B_{\epsilon_x}(x) \cap E) = A$.

Next, assume that A is a union of open balls $\bigcup_{x \in J} (B_{\epsilon_x}(x) \cap E)$ in E (where J is an arbitrary indexing set). Then $A = (\bigcup_{x \in J} B_{\epsilon_x}(x)) \cap E$, which is open in E since we know that $\bigcup_{x \in J} B_{\epsilon_x}(x)$ is open by theorem 10.21. □

Note that if $E = X$ then the preceding theorem states that a set is closed if it contains all of its limit points, and a set is open if and only if it is a union of open balls.

Definition 73

Let A be a set in a metric space X and let A' be the set of limit points of A. The *closure* of A, denoted $\overline{A} = A \cup A'$. The *interior* A° of A is the set of all points p of A so that $B_\epsilon(p) \subset A$ for some $\epsilon > 0$. The *boundary* $\partial(A)$ of A is the set of all points p of X so that $B_\epsilon(p)$ contains a point of A and a point of $X \setminus A$ for each $\epsilon > 0$.

Theorem 10.26. *Let $A \subseteq X$, where (X, d) is a metric space, and let $p \in X$. Then $p \in \partial(A)$ if and only if every open set containing p contains a point of A and a point of the complement of A.*

Proof. Assume $p \in \partial(A)$. Let U be an open set containing p. For some $\epsilon > 0$ we know that $B_\epsilon(p) \subseteq U$. Since $p \in \partial(A)$ we know that $B_\epsilon(p)$ contains a point of A and a point of the complement of A, and thus U does as well.

Assume that every open set containing p contains a point of A and a point of the complement of A. Let $\epsilon > 0$. Since we know $B_\epsilon(p)$ is open, it follows that $B_\epsilon(p)$ contains a point of A and a point of the complement of A, which means that $p \in \partial(A)$. □

Theorem 10.27. *Let E be a subset of a metric space X. Then the closure of E is the intersection of all closed sets containing E and the interior of E is the union of all open sets contained in E.*

Proof. Let A be a closed set containing E. Then A contains all of its limit points. All limit points of E are limit points of A by Exercise 10.5, so A contains all limit points of E and therefore $\overline{E} \subseteq A$, so $\overline{E}$ is a subset of the intersection of all closed sets containing E. If p is a limit point of $\overline{E}$ then every open set U containing p contains a point q of $\overline{E}$ distinct from p. If q is not an element of E then q is a limit point of E, so U contains infinitely many points of E and therefore U contains points of E distinct from p by Exercise 10.13. Hence, p is a limit point of E and therefore an element of $\overline{E}$. Hence $\overline{E}$ contains all of its limit points and is closed. Thus, the intersection of all closed sets containing E is a subset of $\overline{E}$, and is thus equal to $\overline{E}$.

By definition, for each $x \in E^\circ$ we can find $\epsilon_x > 0$ so that $B_{\epsilon_x}(x) \subseteq E$. Since $B_{\epsilon_x}(x)$ is open we know that if $y \in B_{\epsilon_x}(x)$ then $y \in B_\delta(y) \subseteq B_{\epsilon_x}(x) \subseteq E$, which means $y \in E^\circ$. Hence, $B_\epsilon(x) \subseteq E^\circ$. Thus, $E^\circ = \bigcup_{x \in E^\circ} B_{\epsilon_x}(x)$, which is open. Hence, E° is a subset of the union of all open sets which are contained in E. If V is an open set which is contained in E then for every $p \in V$ we can find $\gamma > 0$ so that $B_\gamma(p) \subseteq V \subseteq E$. Thus, $V \subseteq E^\circ$. It follows that the union of all open sets contained in E is a subset of E° and is therefore equal to E°.

□

Theorem 10.28. *Let $E \subseteq X$, a metric space. Then $\partial(E) = \overline{E} \setminus E^\circ$.*

Proof. Let $x \in \partial(E)$. Then for each $\epsilon > 0$ we know that $B_\epsilon(x) \not\subseteq E$ so no open subset of E contains x and $x \notin E^\circ$. If $x \in E$ then $x \in \overline{E}$. If $x \notin E$ then x is a limit point of E since every open ball containing x contains a point of E. Hence, $x \in \overline{E}$. Thus, $\partial(E) \subseteq \overline{E} \setminus E^\circ$.

Let $x \in \overline{E} \setminus E^\circ$. Since $x \notin E^\circ$, for each $\epsilon > 0$ the set $B_\epsilon(x) \not\subseteq E$, which means that $B_\epsilon(x)$ contains a point of the complement of E. Since $x \in \overline{E}$, either $x \in E$ or x is a limit point of E. If $x \in E$ then every open ball about x contains a point of E (namely x). If x is a limit point of E then every open ball about x contains a point of E by definition of limit point. Thus, every open ball about x contains a point of E and a point not in E, which means $x \in \partial(E)$. □

Theorem 10.29. *Let $A, B \subseteq \mathbb{R}^n$. Then:*
(a) $\overline{A \cup B} = \overline{A} \cup \overline{B}$
(b) $\overline{A \cap B} \subseteq \overline{A} \cap \overline{B}$
(c) $(A \cap B)^\circ = A^\circ \cap B^\circ$
(d) $(A \cup B)^\circ \supseteq A^\circ \cup B^\circ$

Proof. (a) Let $\mathbf{x} \in \overline{A \cup B}$. Then either $\mathbf{x} \in A \cup B$ or $\mathbf{x}$ is a limit point of $A \cup B$. If $\mathbf{x} \in (A \cup B)$ then $\mathbf{x} \in \overline{A} \cup \overline{B}$. If $\mathbf{x}$ is a limit point of $A \cup B$ then either $\mathbf{x}$ is a limit point of A or $\mathbf{x}$ is a limit point of B, so $\mathbf{x} \in \overline{A}$ or $\mathbf{x} \in \overline{B}$, so $\mathbf{x} \in \overline{A} \cup \overline{B}$.

Let $\mathbf{x} \in \overline{A} \cup \overline{B}$. Then either $\mathbf{x} \in \overline{A}$ or $\mathbf{x} \in \overline{B}$. Without loss of generality we may assume $\mathbf{x} \in \overline{A}$. Thus, either $\mathbf{x} \in A$ or $\mathbf{x}$ is a limit point of A. If $\mathbf{x} \in A$ then $\mathbf{x} \in \overline{A \cup B}$. If $\mathbf{x}$ is a limit point of A then $\mathbf{x}$ is a limit point of $A \cup B$. Thus, $\mathbf{x} \in \overline{A \cup B}$. Hence, $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

(b) Let $\mathbf{x} \in \overline{A \cap B}$. Either $\mathbf{x} \in (A \cap B)$ or $\mathbf{x}$ is a limit point of $A \cap B$. If $\mathbf{x} \in A \cap B$ then $\mathbf{x} \in \overline{A} \cap \overline{B}$. If $\mathbf{x}$ is a limit point of $A \cap B$ then $\mathbf{x}$ is a limit point of both A and B, so $\mathbf{x} \in \overline{A}$ and $x \in \overline{B}$ and thus $\mathbf{x} \in \overline{A} \cap \overline{B}$. Hence, $\overline{A \cap B} \subseteq \overline{A} \cap \overline{B}$.

(c) Let $\mathbf{x} \in (A \cap B)^\circ$. Then there is an open set U so that $\mathbf{x} \in U \subseteq (A \cap B)$. Since $U \subseteq A$ and $U \subseteq B$ it follows that $\mathbf{x} \in A^\circ$ and $\mathbf{x} \in B^\circ$, so $\mathbf{x} \in A^\circ \cap B^\circ$.

Let $\mathbf{x} \in A^\circ \cap B^\circ$. Then there are open sets V_1, V_2 so that $\mathbf{x} \in V_1 \subseteq A$ and $\mathbf{x} \in V_2 \subseteq B$, which means that $\mathbf{x} \in (V_1 \cap V_2) \subseteq (A \cap B)$, which means that $\mathbf{x} \in (A \cap B)^\circ$. Hence, $(A \cap B)^\circ = A^\circ \cap B^\circ$.

(d) Let $\mathbf{x} \in A^\circ \cup B^\circ$. Then either $\mathbf{x}$ is in an open set contained in A or in an open set contained in B. Withoug loss of generality, we may assume there is an open set U so that $x \in U \subseteq A \subseteq (A \cup B)$. Thus, $\mathbf{x} \in (A \cup B)^\circ$ and $(A \cup B)^\circ \supseteq A^\circ \cup B^\circ$.

□

Theorem 10.30. *Open Set Characterization of Continuity. Let (X, d) and (Y, ρ) be metric spaces and let $E \subset X$. Let $f : E \to Y$. Then:*

(a) (Local Form): The function f is continuous at a point $p \in E$ if and only if for every open set V containing $f(p)$, there is an open set U containing p such that $U \cap E \subseteq f^{-1}(V)$.

(b) (Global Form): The function f is continuous if and only if, for every open subset V of Y, the set $f^{-1}(V)$ is open in E.

Proof. (a) First, assume that f is continuous at $p \in E$. Let V be open in Y containing $f(p)$. Since V is open there is an $\epsilon > 0$ so that $B_\epsilon(f(p)) \subseteq V$. Since f is continuous there is a $\delta > 0$ so that if $d(p, x) < \delta$ and $x \in E$ then $\rho(f(x), f(p)) < \epsilon$ for all $x \in E$. Thus $B_\delta(p) \cap E \subseteq f^{-1}(V)$.

Next, assume that for every open set V containing $f(p)$, there is an open set U containing p such that $U \cap E \subseteq f^{-1}(V)$. Let $\epsilon > 0$. Since $B_\epsilon(f(p))$ is open, there is an open set U so that $p \in U \cap E \subset f^{-1}(B_\epsilon(f(p))$, which means that there is a $\delta > 0$ so that $B_\delta(p) \cap E \subseteq U \cap E \subseteq f^{-1}(B_\epsilon(f(p))$. Hence, if $d(x, p) < \delta$ and $x \in E$ then $\rho(f(x), f(p)) < \epsilon$, which means that f is continuous at p.

(b) First, assume f is continuous, and let V be open in Y. If $V \cap f(E) = \emptyset$ then $f^{-1}(V)$ is empty and therefore open. Assume $V \cap f(E) \neq \emptyset$. Let $p \in f^{-1}(V)$. By (a) there is an open set U_p in X so that $p \in U_p \cap E \subseteq f^{-1}(V)$. Choosing such an open set U_p for each $p \in f^{-1}(V)$, we see that $U = \displaystyle\bigcup_{p \in f^{-1}(V)} U_p$ is an open set in X so that $U \cap E = f^{-1}(V)$, which is therefore open in E.

Next, assume that the inverse of every open set in Y is open in E. Let $p \in E$. Then for every open set V containing $f(p)$, there is an open set U containing p such that $U \cap E = f^{-1}(V)$, which means that $U \cap E \subseteq f^{-1}(V)$. Hence, by part (a) it follows that f is continuous at p for every $p \in E$, so f is continuous.

□

Definition 74

We say a function $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation if $T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T(\mathbf{x}) + \beta T(\mathbf{y})$ for all $\alpha, \beta \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then the *operator norm* $|T| = \sup\limits_{\mathbf{x} \neq \mathbf{0}} \dfrac{|T(\mathbf{x})|}{|\mathbf{x}|}$. We use the notation $\mathbf{e}_i$ to denote the ith standard basis vector for $\mathbf{R}^n$, which is the vector $\mathbf{e}_i =< 0, 0, ..., 0, 1, 0, 0, ..., 0 >$ where every coordinate entry of the vector is zero except the ith entry, which is one.

Theorem 10.31. *Let $\boldsymbol{x} = (x_1, x_2, x_3, ..., x_n) \in \mathbb{R}^n$. Then $|\boldsymbol{x}| \leq \displaystyle\sum_{i=1}^{n} |x_i|$.*

Proof. We note that $\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{e}_i$, so by the Triangle Inequality (part (c)) we have that $|\mathbf{x}| = |\sum_{i=1}^{n} x_i\mathbf{e}_i| \le \sum_{i=1}^{n} |x_i\mathbf{e}_i| = \sum_{i=1}^{n} |x_i|$. □

Theorem 10.32. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then the operator norm $|T|$ exists and is a non-negative real number. If the matrix for T is A with rows $A_1, A_2, ..., A_m$ then $|T|$ is less than or equal to $\sqrt{m}\sum_{i=1}^{m}|A_i| \le \sqrt{m}\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$.*

Proof. We can find a matrix $A = [a_{ij}]_{m\times n}$ so that $T(\mathbf{x}) = A\mathbf{x}$ by Theorem 14.10. Thus, if A_i is the ith row vector of A then $A\mathbf{x} = [A_i \cdot \mathbf{x}]_{1\times m}$. Let $C = \max\{|A_1|, ..., |A_m|\}$. Since $|A_i \cdot \mathbf{x}| \le |A_i||\mathbf{x}|$ for each i by the Cauchy Schwarz inequality, we know that $|T(\mathbf{x})| = \sqrt{\sum_{i=1}^{m}(A_i \cdot \mathbf{x})^2} \le \sqrt{\sum_{i=1}^{m}|A_i|^2|\mathbf{x}|^2} \le C\sqrt{\sum_{i=1}^{m}|\mathbf{x}|^2} = C\sqrt{m}|\mathbf{x}|$. Hence, $\sup_{\mathbf{x}\neq\mathbf{0}}\frac{|T(\mathbf{x})|}{|\mathbf{x}|} \le C\sqrt{m} \le \sqrt{m}\sum_{i=1}^{m}|A_i| \le \sqrt{m}\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$ by Theorem 10.31. □

Definition 75

A collection $\mathcal{C}$ of sets is a *cover* of a set $A \subseteq X$ if the union of $\mathcal{C}$ contains A. If the sets in $\mathcal{C}$ are open then we call $\mathcal{C}$ an *open cover* or open covering of A. We say that a set $A \subseteq X$ is *compact* if every open cover $\mathcal{C}$ of A has a finite subset F which is also a cover of A. We call F a finite *subcover* of A.

Theorem 10.33. *Let K be a compact subset of a metric space X and let $f : K \to Y$ be continuous, where Y is a metric space. Then $f(K)$ is compact.*

Proof. Let $C = \{U_\alpha\}_{\alpha\in J}$ be an open cover of $f(K)$. For each $\alpha \in J$ choose V_α open in X so that $V_\alpha \cap K = f^{-1}(U_\alpha)$. Then the collection of V_α sets covers K and has a finite subcover $V_{\alpha_1}, V_{\alpha_2}, .., V_{\alpha_t}$. The corresponding open sets $U_{\alpha_1}, U_{\alpha_2}, .., U_{\alpha_t}$ are a finite subset of C which covers $f(K)$. Thus $f(K)$ is compact. □

Theorem 10.34. *Bolzano-Weierstrass Theorem in $\mathbb{R}^m$.*

(a) Let $\{\boldsymbol{x}_n\}$ be a sequence of points in $\mathbb{R}^m$ which is bounded. Then $\{\boldsymbol{x}_n\}$ has a convergent subsequence.

(b) Every bounded infinite set in $\mathbb{R}^m$ has a limit point.

Proof. (a) Let x_n^i denote the ith component of $\mathbf{x}_n$. Since $\{\mathbf{x}_n\}$ is bounded, it follows that each $\{x_n^i\}$ is bounded. By the Bolzano-Weierstrass Theorem in $\mathbb{R}$ we know that $\{x_n^1\}$ has a convergent subsequence $\{x^1_{n_{(j_1)}}\}_{\{j_1\in\mathbb{N}\}} \to p_1$. Likewise, $\{x^2_{n_{(j_1)}}\}$ has a convergent subsequence $\{x^2_{n_{(j_1,j_2)}}\} \to p_2$. We continue in this manner until we pick a subsequence $\{x^m_{n_{(j_1,j_2,\ldots,j_m)}}\} \to p_m$. Then $\{\mathbf{x}_{n_{(j_1,j_2,\ldots,j_m)}}\} \to \mathbf{p} = (p_1, p_2, \ldots, p_m)$.

(b) Let E be a bounded infinite set in $\mathbb{R}^m$. We choose a sequence of distinct points of E inductively by choosing $\mathbf{x}_1 \in E$ and then if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ have been chosen to be distinct points of E for some $k \in \mathbb{N}$, then since E is infinite there are points of E which have not been chosen so we can pick $\mathbf{x}_{k+1} \in E \setminus \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_k\}$. Then the sequence $\{\mathbf{x}_n\}$ is a sequence of distinct points which is bounded and therefore has a subsequence $\{\mathbf{x}_{n_i}\}$ which converges to a point $\mathbf{p}$ by (a). For every $\epsilon > 0$, this means that there is a $t \in \mathbb{N}$ so that if $i \geq t$ then $\mathbf{x}_{n_i} \in B_\epsilon(\mathbf{p})$, which means that $B_\epsilon(\mathbf{p})$ contains infinitely many points of E and so $\mathbf{p}$ is a limit point of E.

□

Theorem 10.35. *Let K be a closed and bounded set in $\mathbb{R}^n$ and let $f : K \to \mathbb{R}^m$ be continuous. Then f is uniformly continuous.*

Proof. Suppose f is not uniformly continuous. Then there is an $\epsilon > 0$ so that for every $\delta > 0$ there are points $\mathbf{x}, \mathbf{y} \in K$ so that $|\mathbf{x} - \mathbf{y}| < \delta$ and $|f(\mathbf{x}) - f(\mathbf{y})| \geq \epsilon$. For each $n \in \mathbb{N}$ choose points $\mathbf{x}_n, \mathbf{y}_n \in K$ so that $|\mathbf{x}_n - \mathbf{y}_n| < \dfrac{1}{n}$ and $|f(\mathbf{x}_n) - f(\mathbf{y}_n)| \geq \epsilon$. By the Bolzano Weierstrass Theorem, we can find a convergent subsequence $\{\mathbf{x}_{n_i}\} \to \mathbf{p}$ where $\mathbf{p} \in K$ since K is closed. Since $\{|\mathbf{x}_{n_i} - \mathbf{y}_{n_i}|\} \to 0$ we also know that $\{\mathbf{y}_{n_i}\} \to \mathbf{p}$. Since f is continuous it follows that $\{f(\mathbf{x}_{n_i})\} \to f(\mathbf{p})$ and $\{f(\mathbf{y}_{n_i})\} \to f(\mathbf{p})$, and hence $\{|f(\mathbf{x}_{n_i}) - f(\mathbf{y}_{n_i})|\} \to 0$, which is impossible since $|f(\mathbf{x}_{n_i}) - f(\mathbf{y}_{n_i})| \geq \epsilon$ for all $i \in \mathbb{N}$. Hence, f is uniformly continuous. □

Definition 76

Let X be a metric space. We say that D is *dense* in X if every open ball in X contains a point of D. We say X is *separable* if there is a countable set $D \subseteq X$ which is dense in X. We say that a collection $\mathcal{B}$ of non-empty open sets is a *basis* for X if for every $x \in X$ and open set U containing p there is some $B \in \mathcal{B}$ so that $p \in B \subset U$. We say that a set $E \subseteq X$ is *Lindelof* if every open covering of E has a countable subset which is a cover of E (a countable subcover).

Note that by theorem 10.25, the epsilon balls about points of X form a basis for a metric space X.

Theorem 10.36. *A metric space X is separable if and only if X has a countable basis.*

Proof. Let X have countable basis $\mathcal{B} = \{B_n\}_{n \in \mathbb{N}}$. Choose one point p_n from each B_n to form a set $D = \{p_1, p_2, p_3, \ldots\}$. Each open set U contains a set B_n containing a point $p_n \in D$, so D is dense in X.

Let X be separable with countable dense set $D = \{p_1, p_2, p_3, \ldots\}$ and order the positive rational numbers as $\{q_1, q_2, q_3, \ldots\}$. Let $B(i, j) = B_{q_i}(p_j)$ for each pair of positive integers

i, j. Then let $\mathcal{B} = \{B(i,j)\}|i, j \in \mathbb{N}\}$. Since $\mathcal{B}$ is a countable union of countable sets it is countable. Let U be any open set and $p \in U$. Then there is some $\epsilon > 0$ so that $B_{3\epsilon}(p) \subseteq U$ and some $p_n \in B_\epsilon(p)$. Let $\epsilon < q_m < 2\epsilon$ where q_m is rational. Then $p \in B_{q_m}(p_n) \subseteq B_{3\epsilon}(p) \subseteq U$. Thus, $\mathcal{B}$ is a basis for X.

□

Theorem 10.37. *$\mathbb{R}^n$ is separable and has a countable basis.*

Proof. Let $D = \{\mathbf{x} \in \mathbb{R}^n | x_i \in \mathbb{Q}$ for each $1 \le i \le n\}$, the set of all points each of whose coordinates is rational. Let U be any non-empty open set and choose $B_\epsilon(\mathbf{p}) \subseteq U$, where $\mathbf{p} = (p_1, p_2, ..., p_n)$. Then $\displaystyle\prod_{i=1}^{n}(p_i - \frac{\epsilon}{\sqrt{n}}, p_i + \frac{\epsilon}{\sqrt{n}}) \subset B_\epsilon(\mathbf{p})$, and for each $1 \le i \le n$ we can pick a rational number $q_i \in (p_i - \frac{\epsilon}{\sqrt{n}}, p_i + \frac{\epsilon}{\sqrt{n}})$, which means that $\mathbf{q} = (q_1, q_2, q_3, ..., q_n) \in U$ so D is dense in U. Thus, by theorem 10.36, we know that $\mathbb{R}^n$ has a countable basis. □

Theorem 10.38. *Let $E \subset \mathbb{R}^n$. Then E is Lindelof.*

Proof. Let $\mathcal{C}$ be an open cover of E. Let $\mathcal{B}$ be a countable basis for $\mathbb{R}^n$ and let $B = \{B_1, B_2, B_3, ...\}$ be the elements of $\mathcal{B}$ which contain a point of E and are also contained in an element of $\mathcal{C}$. For each $B_i \in B$ we choose some $U_i \in \mathcal{C}$ which contains B_i, and let $C = \{U_1, U_2, U_3, ...\}$. For each $p \in E$ there is some $U \in \mathcal{C}$ which contains p and thus some $B_j \in B$ so that $p \in B_j \subseteq U$ which means that $p \in U_j$. Hence, C is a countable subset of $\mathcal{C}$ which covers E. □

Theorem 10.39. *Heine-Borel Theorem. Let K be a subset of $\mathbb{R}^m$. Then K is compact if and only if K is closed and bounded.*

Proof. First, assume that K is compact. Then $\{B_n(\mathbf{0})\}_{n\in\mathbb{N}}$ covers K so it has a finite subcover $F = \{B_{n_1}(\mathbf{0}), B_{n_2}(\mathbf{0}), ..., B_{n_j}(\mathbf{0})\}$ where $n_1 < n_2 < ... < n_j$, so $\bigcup F = B_{n_j}(\mathbf{0})$ contains K, and K is bounded.

Next, let $\mathbf{p}$ be a limit point of K and suppose that $\mathbf{p} \notin K$. Then if we set $U_n = \mathbb{R}^m \setminus \overline{B_{\frac{1}{n}}(\mathbf{p})}$ then $\{U_n\}_{n\in\mathbb{N}}$ is an open cover of K and has a finite subcover $F = \{\mathbb{R}^m \setminus \overline{B_{\frac{1}{n_1}}(\mathbf{p})}, \mathbb{R}^m \setminus \overline{B_{\frac{1}{n_2}}(\mathbf{p})}, ..., \mathbb{R}^m \setminus \overline{B_{\frac{1}{n_j}}(\mathbf{p})}\}$ where $n_1 < n_2 < ... < n_j$, so $\bigcup F = \mathbb{R}^m \setminus \overline{B_{\frac{1}{n_j}}(\mathbf{p})}$ contains K. But this is impossible because $\mathbf{p}$ is a limit point of K so $B_{\frac{1}{n_j}}(\mathbf{p})$ must contain points of K.

Finally, assume that K is closed and bounded. Let $\mathcal{C}$ be an open cover of K. Suppose $\mathcal{C}$ has no finite subcover. By theorem 10.38 there is a countable subcover $\{U_n\}_{n\in\mathbb{N}}$. For each $n \in \mathbb{N}$ we choose a point $\mathbf{p}_n \in K \setminus \displaystyle\bigcup_{j=1}^{n} U_j$. Since K is bounded, by the Bolzano-Weierstrass Theorem there is a convergent subsequence $\{\mathbf{p}_{n_i}\} \to \mathbf{p}$, where $\mathbf{p} \in K$ since K is closed. Thus, there is some t so that $\mathbf{p} \in U_t$ which means that there is a $k \in \mathbb{N}$ so that if $i \ge k$ then $\mathbf{p}_{n_i} \in U_t$, which we know is false if $i \ge t$, a contradiction. Hence, we conclude that K is compact.

□

Theorem 10.40. *Let K be a compact subset of $\mathbb{R}$. Then K has a first point and a last point.*

Proof. We know that K is closed and bounded by the Heine-Borel Theorem, which means that K has a least upper bound u and a greatest lower bound b. Suppose that $u \notin K$. Then by the Approximation Property, for every $\epsilon > 0$ there is some $x \in K$ so that $u - \epsilon < x < u$, which means that u is a limit point of K. But this means that $u \in K$ since K contains all of its limit points. Likewise, $b \in K$. □

Theorem 10.41. *The Extreme Value Theorem (for metric spaces). Let X be a metric space and let $f : K \to \mathbb{R}$ be continuous, where K is a compact subset of X. Then there are points $s, t \in K$ so that $f(s) \leq f(x) \leq f(t)$ for all $x \in K$.*

Proof. By Theorem 10.33, we know that $f(K)$ is compact and therefore has a first and last point. □

Theorem 10.42. *The Lebesgue Number Lemma. Let $\mathcal{C}$ be an open cover of a compact set K in a metric space X. Then there is a number $\delta > 0$ so that for any $x \in K$ the ball $B_\delta(x)$ is a subset of an element of $\mathcal{C}$. Furthermore, if S is a set with diameter less than δ and $S \cap K \neq \emptyset$ then S is a subset of some element of $\mathcal{C}$.*

Proof. For each $x \in K$, we choose an $\epsilon_x > 0$ so that $B_{2\epsilon_x}(x) \subset W$ for some $W \in \mathcal{C}$. Then $\{B_{\epsilon_x}(x) | x \in K\}$ is an open cover of K, and since K is a compact space we know that there is a finite subcover $F = \{B_{\epsilon_{x_1}}(x_1), B_{\epsilon_{x_2}}(x_2), B_{\epsilon_{x_3}}(x_3), ...B_{\epsilon_{x_n}}(x_n)\}$. Let $\delta = \min\{\epsilon_{x_1}, \epsilon_{x_2}, ..., \epsilon_{x_n}\}$.

Let $x \in K$. Then for some x_i we know that $x \in B_{\epsilon_{x_i}}(x_i)$, so if $y \in B_\delta(x)$ then $\rho(y, x_i) < 2\epsilon_{x_i}$, so $B_\delta(x) \subseteq B_{2\epsilon_{x_i}}(x_i) \subseteq W$ for some $W \in \mathcal{C}$.

If $diam(S) < \delta$ and $x \in K \cap S$ then $S \subset B_\delta(x)$ which is a subset of an element of $\mathcal{C}$. □

Definition 77

Let $E \subseteq \mathbb{R}^n$. A pair of non-empty sets H and K is a *separation* of E if $H, K \subseteq E$, $H \cap \overline{K} = \emptyset = K \cap \overline{H}$ and $H \cup K = E$. We say $E \subseteq \mathbb{R}^n$ is *connected* if it has no separation. We say that E is *path connected* if, for each pair of points $\mathbf{p}, \mathbf{q} \in E$ there is a continuous function $f : [a, b] \to E$ so that $f(a) = \mathbf{p}$ and $f(b) = \mathbf{q}$. We say that E is *polygonally connected* if, for each pair of points $\mathbf{p}, \mathbf{q} \in E$ there is a finite sequence of line segments $L(\mathbf{x}_1, \mathbf{x}_2), L(\mathbf{x}_2, \mathbf{x}_3), ..., L(\mathbf{x}_{m-1}, \mathbf{x}_m)$ which are contained in E so that $\mathbf{x}_1 = \mathbf{p}$, and $\mathbf{x}_m = \mathbf{q}$. We refer to the union of these line segments as a *polygonal path* from $\mathbf{p}$ to $\mathbf{q}$. If $L(\mathbf{p}, \mathbf{q}) \subseteq E$ for each $\mathbf{p}, \mathbf{q} \in E$ then E is *convex*.

Theorem 10.43. *Let $E \subseteq \mathbb{R}^n$. Let H and K be disjoint non-empty subsets of E so that $H \cup K = E$. Then H and K are a separation of E if and only if H and K are open in E.*

Proof. First, assume that H and K are a separation of E. Then H contains no limit points of K, so for every point $\mathbf{p} \in H$ we can find $\epsilon_{\mathbf{p}} > 0$ so that $B_{\epsilon_{\mathbf{p}}}(\mathbf{p}) \cap K = \emptyset$. Thus, $U = \bigcup\limits_{\mathbf{p} \in H} B_{\epsilon_{\mathbf{p}}}(\mathbf{p})$ is an open set and $U \cap E = H$, which means that H is open in E. Likewise, K is open in E.

Next, assume that H and K are open in E. We already know that H and K are disjoint non-empty subsets of E so that $H \cup K = E$, so we need only verify that H contains no limit points of K and K contains no limit points of H. Let $\mathbf{p} \in H$. There is an open set V so that $V \cap E = H$ since H is open in E. Hence, $\mathbf{p}$ is contained in an open set which does not intersect K, so $\mathbf{p}$ is not a limit point of K. Similarly, K contains no limit points of H. □

Theorem 10.44. *Let C be a connected subset of $\mathbb{R}^n$ and let $f : C \to \mathbb{R}^m$ be continuous. Then $f(C)$ is connected.*

Proof. Suppose $f(C)$ is not connected. Then it has a separation H and K, where $H = U \cap f(C)$ and $K = V \cap f(C)$ for some open sets U and V in $\mathbb{R}^m$, where H and K are non-empty, disjoint and have $H \cup K = f(C)$. By theorem 10.30, this means that $f^{-1}(H)$ and $f^{-1}(K)$ are open in C. Since they are also disjoint and non-empty and have union equal to C, it follows that C is not connected, a contradiction. □

Theorem 10.45. *Intermediate Value Theorem in $\mathbb{R}^n$. Let H be a connected subset of $\mathbb{R}^n$ and let $f : H \to \mathbb{R}$ be continuous, where $f(\boldsymbol{a}) < k < f(\boldsymbol{b})$ for some $\boldsymbol{a}, \boldsymbol{b} \in H$. Then there is some $\boldsymbol{c} \in H$ so that $f(\boldsymbol{c}) = k$.*

Proof. From theorem 10.44 we know that $f(H)$ is connected and is therefore an interval by theorem 7.51. Thus, by the definition of interval, $k \in f(H)$. □

Theorem 10.46. *Every polygonally connected set is path connected and every path connected set is connected.*

Proof. Let E be polygonally connected. Let $\mathbf{p}, \mathbf{q} \in E$. Then there are line segments $L(\mathbf{x}_0, \mathbf{x}_2), L(\mathbf{x}_2, \mathbf{x}_3), ..., L(\mathbf{x}_{m-1}, \mathbf{x}_m) \subseteq E$ where $\mathbf{x}_0 = \mathbf{p}$, and $\mathbf{x}_m = \mathbf{q}$, whose union is a polygonal path $P(\mathbf{p}, \mathbf{q})$. Define $f : [0, m+1]$ by $f(x) = (i+1-x)\mathbf{x}_i + (x-i)\mathbf{x}_{i+1}$ for all $i \le x \le i+1$, for each $0 \le i \le m-1$. Then f is continuous function whose image is the polygonal path $P(\mathbf{p}, \mathbf{q})$. Thus, E is path connected.

Assume that E is path connected. Suppose E is not connected. Then E has a separation consisting of sets H and K which are non-empty, neither of which contains a limit point or point of the other, and whose union is E. Let $\mathbf{p} \in H$ and $\mathbf{q} \in K$. Since E is path connected there is a continuous function $f : [a, b] \to E$ so that $f(a) = \mathbf{p}$ and $f(b) = \mathbf{q}$. Since $[a, b]$ is connected by Theorem 7.51 and f is continuous, we know that $f([a, b]) = C$ is a connected set by Theorem 10.44.

Since H and K are a separation of E, we know that H and K are open in E, so there are open sets U and V so that $U \cap E = H$ and $V \cap E = K$. Let $H' = H \cap C = U \cap C$ and $K' = K \cap C = V \cap C$. Then H' and K' are open in C and are a separation of C, which contradicts the fact that C is connected. We conclude that E is connected.

□

Theorem 10.47. *Let U be a connected open subset of $\mathbb{R}^n$. Then U is polygonally connected.*

Proof. Let $\mathbf{a} \in U$. Let $S = \{\mathbf{x} \in U|$ there is a polygonal path $P(\mathbf{a}, \mathbf{x}) = L(\mathbf{x}_0, \mathbf{x}_1) \cup L(\mathbf{x}_1, \mathbf{x}_2) \cup ... \cup L(\mathbf{x}_{m-1}, \mathbf{x}_m) \subset U$ from $\mathbf{a} = \mathbf{x}_0$ to $\mathbf{x} = \mathbf{x}_m\}$. We will prove that S is open, and that S and closed in U.

To show S is open, let $\mathbf{p} \in S$. Then $\mathbf{p} \in U$ so we can find $\epsilon > 0$ so that $B_\epsilon(\mathbf{p}) \subseteq U$. Let $\mathbf{z} \in B_\epsilon(\mathbf{p})$. Since $\mathbf{p} \in S$ there is a polygonal path $P(\mathbf{a}, \mathbf{p}) = L(\mathbf{a}, \mathbf{x}_1) \cup L(\mathbf{x}_1, \mathbf{x}_2) \cup ... \cup L(\mathbf{x}_{m-1}, \mathbf{p}) \subset U$. Thus, $L(\mathbf{a}, \mathbf{x}_1) \cup L(\mathbf{x}_1, \mathbf{x}_2) \cup ... \cup L(\mathbf{x}_{m-1}, \mathbf{p}) \cup L(\mathbf{p}, \mathbf{z})$ is a polygonal path from $\mathbf{a}$ to $\mathbf{z}$ which is contained in U, which means that $B_\epsilon(\mathbf{p}) \subseteq U$ and therefore S is open.

To show that S contains all limit points of S which are contained in U, let $\mathbf{q} \in U$ be a limit point of S. Let $\gamma > 0$ so that $B_\gamma(\mathbf{q}) \subset U$. Then $B_\gamma(\mathbf{q})$ contains a point $\mathbf{w} \in S$. Since $\mathbf{w} \in S$, there is a polygonal path $P(\mathbf{a}, \mathbf{w}) = L(\mathbf{a}, \mathbf{x}_1) \cup L(\mathbf{x}_1, \mathbf{x}_2) \cup ... \cup L(\mathbf{x}_{m-1}, \mathbf{w}) \subset U$. But then $P(\mathbf{a}, \mathbf{w}) \cup L(\mathbf{w}, \mathbf{q})$ is a polygonal path from $\mathbf{a}$ to $\mathbf{q}$ which is contained in U, meaning that $\mathbf{q} \in U$.

Either $S = U$, in which case U is polygonally connected, or not. If not then let $H = S$ and let $K = U \setminus S$. Then H, K are disjoint, non-empty and have union equal to U, and since H is open we know that H contains no limit points of K, and since H contains all of its limit points which are in U, it follows that K contains no limit points of H. Hence, H and K are a separation of U, which is impossible since U is connected. We conclude that $S = U$ and U is polygonally connected. □

Theorem 10.48. *Let $f : [a, b] \to \mathbb{R}$ and let $G = \{(x, f(x)) \in \mathbb{R}^2 | x \in [a, b]\}$ be the graph of f. Then f is continuous if and only if G is both closed and connected.*

Proof. First, assuming f is continuous we know that $F(x) = (x, f(x))$ is continuous by Theorem 10.17, which means that $F([a, b]) = G$ is connected and compact and therefore closed (by Theorems 10.44, 10.33 and the Heine-Borel Theorem).

Next, assume that G is connected and closed. We first check that f satisfies the property that if $a \leq c < d \leq b$ and $f(c) < r < f(d)$ or $f(c) > r > f(d)$ then $f(q) = r$ for some $q \in (c, d)$. Suppose $f(c) < r < f(d)$ and there is no $q \in (c, d)$ so that $f(q) = r$. Then $H = \{(x, y) \in G | x < c \text{ or } x < d \text{ and } y < r\}$ and $K = \{(x, y) \in G | x > d \text{ or } x > c \text{ and } y > r\}$ is a separation of G, contradicting that G is connected. In the case that $f(c) > r > f(d)$ we can negate f and apply the previous case. We will refer to this condition as the intermediate value property.

Suppose that f has a discontinuity at a point $c \in [a, b]$. Then we can find $\epsilon > 0$ so that for every $\delta > 0$ we can find $x \in [a, b]$ so that $|x - c| < \delta$ but $|f(x) - f(c)| \geq \epsilon$. For each

$n \in \mathbb{N}$ choose $x_n \in [a,b]$ so that $|x_n - c| < \dfrac{1}{n}$ and $|f(x_n) - f(c)| \geq \epsilon$. Either there are infinitely many integers n so that $f(x_n) \geq f(c) + \epsilon$ or there are infinitely many integers n so that $f(x_n) \leq f(c) - \epsilon$. Assume there are infinitely many integers n so that $f(x_n) \geq f(c) + \epsilon$, yielding a subsequence $\{n_i\}$ so that $f(x_{n_i}) \geq f(c) + \epsilon$ for each $i \in \mathbb{N}$.

Since f has the intermediate value property, for each $i \in \mathbb{N}$ we can find a point c_i so $|c_i - c| < |x_{n_i} - c|$ so that $f(c_i) = c + \epsilon$. Then $\{(c_i, f(c) + \epsilon)\} \to (c, f(c) + \epsilon)$, which means that $(c, f(c) + \epsilon)$ is a limit point of G, so G is not closed (since G contains only the point $(c, f(c))$ on the line $x = c$). The argument is similar if there are infinitely many integers n so that $f(x_n) \leq f(c) - \epsilon$. This contradiction implies that f is continuous.

□

Exercises:

Exercise 10.1. *Let $E \subseteq \mathbb{R}^n$ and let $p \notin E$. Then $p \in \partial(E)$ if and only if p is a limit point of E.*

Exercise 10.2. *Let $U \subseteq \mathbb{R}^n$. Then U is open if and only if U contains none of its boundary points.*

Exercise 10.3. *Let $A \subseteq \mathbb{R}^n$. Then A is closed if and only if A contains all of its boundary points.*

Exercise 10.4. *If p is a limit point of $\displaystyle\bigcup_{i=1}^{n} E_i$ in metric space (X, d) then p is a limit point of E_k for some k.*

Exercise 10.5. *If p is a limit point of A and $A \subset B$ in a metric space X then p is a limit point of B.*

Exercise 10.6. *A point p is a limit point of a set E in a metric space (X, d) if and only if every open set containing p contains infinitely many points of E, which is true if and only if there is a sequence of points in $E \setminus \{p\}$ which converges to p.*

Exercise 10.7. *If E_1, E_2 are connected subsets of $\mathbb{R}^n$ which share a common point $\boldsymbol{p}$ then $E_1 \cup E_2$ is connected.*

Exercise 10.8. *Give examples of open sets $U_1, U_2, U_3, \ldots$ so that $\displaystyle\bigcap_{i=1}^{\infty} U_i$ is not open.*

Exercise 10.9. *Show that no subset of $\mathbb{Q}^n$ containing more than one point is connected.*

Exercise 10.10. *The empty set is the only proper subset of $\mathbb{R}^n$ which is both closed and open.*

Exercise 10.11. *Let $E \subseteq \mathbb{R}^n$. Then $\overline{E} = E \cup \partial(E)$.*

Exercise 10.12. *If $A \subseteq K \subseteq \mathbb{R}^n$ and K is closed then A is closed in K if and only if A is closed.*

Exercise 10.13. *A point p of a metric space (X, d) is a limit point of a set $E \subseteq X$ if and only if every open set containing p contains infinitely many points of E.*

Solutions:

Solution to Exercise 10.1. *Let $E \subseteq \mathbb{R}^n$ and let $p \notin E$. Then $p \in \partial(E)$ if and only if p is a limit point of E.*

Proof. Every open set containing $\mathbf{p}$ contains a point not in E (namely $\mathbf{p}$). Thus, $\mathbf{p}$ is a limit point of E if and only if every open set containing $\mathbf{p}$ contains a point of E (since such a point will always be distinct from $\mathbf{p}$), which is also true if and only if $\mathbf{p}$ is a boundary point of E.

□

Solution to Exercise 10.2. *Let $U \subseteq \mathbb{R}^n$. Then U is open if and only if U contains none of its boundary points.*

Proof. Assume U is open. Let $\mathbf{p} \in U$. Then since U is open there is an open ball containing $\mathbf{p}$ which is contained in U and therefore contains no points which are not contained in U, so $\mathbf{p}$ is not a boundary point of U.

Assume that U contains none of its boundary points. Let $\mathbf{p} \in U$. Since $\mathbf{p} \in U$ we know $\mathbf{p}$ is not a boundary point of U, which means that there is an open ball containing $\mathbf{p}$ which contains no points which are not in U (since such a ball must contain $\mathbf{p}$ and therefore contains a point of U). Hence, $\mathbf{p}$ is contained in an open set which is contained in U, which means that U is open. □

Solution to Exercise 10.3. *Let $A \subseteq \mathbb{R}^n$. Then A is closed if and only if A contains all of its boundary points.*

Proof. First, assume that A is closed. Since it has been established that every boundary point of A which is not in A must be a limit point of A, and also that closed sets contain all of their limit points, it follows that every boundary point of A is a point of A.

Next, assume that A contains all of its boundary points. Then if $\mathbf{p} \notin A$ it follows that $\mathbf{p}$ is not a boundary point of A, which means that $\mathbf{p}$ is not a limit point of A since all limit points of A which are not contained in A are boundary points of A. Thus, A contains all of its limit points and is therefore closed. □

Solution to Exercise 10.4. *If p is a limit point of $\displaystyle\bigcup_{i=1}^{n} E_i$ in metric space (X, d) then p is a limit point of E_k for some k.*

Proof. Suppose p is not a limit point of any E_i. For each i we can choose an $\epsilon_i > 0$ so that $B_{\epsilon_i}(p)$ contains no points of E_i distinct from p. Let $\epsilon = \min\{\epsilon_1, \epsilon_2, ..., \epsilon_n\}$. Then $B_\epsilon(p)$ contains no points of $\displaystyle\bigcup_{i=1}^{n} E_i$ distinct from p, a contradiction to p being a limit point of $\displaystyle\bigcup_{i=1}^{n} E_i$. □

Solution to Exercise 10.5. *If p is a limit point of A and $A \subset B$ in a metric space X then p is a limit point of B.*

Proof. If p is a limit point of A then every open set containing p contains a point of A distinct from p, which is also a point of B distinct from p since every point of A is a point of B. Thus, p is a limit point of B.

□

Solution to Exercise 10.6. *A point p is a limit point of a set E in a metric space (X, d) if and only if every open set containing p contains infinitely many points of E, which is true if and only if there is a sequence of points in $E \setminus \{p\}$ which converges to p.*

Proof. By Theorem 10.12, p is a limit point of E if and only if we can find a is a sequence $\{x_n\} \subseteq X \setminus \{p\}$ which converges to p.

Let U be an open set containing p. Suppose there are only finitely many points $p_1, p_2, ..., p_k \in E \setminus \{p\}$. Choose $0 < \epsilon < \min\limits_{1 \leq i \leq k} \{d(p, p_i)\}$ so that $B_\epsilon(p) \subseteq U$. $B_\epsilon(p)$ contains no points of E distinct from p, which is a contradiction.

Next, assume every open set containing p contains infinitely many points of E and let $\delta > 0$. Then the open set $B_\delta(p)$ contains infinitely many points of E and therefore contains points of E distinct from p. Hence, p is a limit point of E.

□

Solution to Exercise 10.7. *If E_1, E_2 are connected subsets of $\mathbb{R}^n$ which share a common point $\mathbf{p}$ then $E_1 \cup E_2$ is connected.*

Proof. Suppose $E_1 \cup E_2$ is not connected and $\mathbf{p} \in E_1 \cap E_2$. Let A, B be a separation of $E_1 \cup E_2$ and let $\mathbf{p} \in A$. Then there are open sets U, V in $\mathbb{R}^n$ so that $A = U \cap (E_1 \cup E_2)$ and $B = V \cap (E_1 \cup E_2)$. If B contains points of E_1 then we know that $U \cap E_1 = B_1$ is non-empty and since A contains $\mathbf{p}$ we know that $U \cap E_1 = A_1$ is non-empty. We also know that $A_1 \cup B_1 = E_1$ and that both A_1 and B_1 are open in E_1 which means that E_1 is not connected, which is impossible. We must conclude that B contains no points of E_1. However, a similar argument shows that B contains no points of E_2, so B is empty, which is a contradiction to the assumption that A and B are a separation of $E_1 \cup E_2$. We conclude that $E_1 \cup E_2$ is connected.

□

Solution to Exercise 10.8. *Give examples of open sets $U_1, U_2, U_3, ...$ so that $\bigcap\limits_{i=1}^{\infty} U_i$ is not open.*

Proof. Let $U_i = B_{\frac{1}{i}}(\mathbf{0})$ for each natural number i. Then $\bigcap\limits_{i=1}^{\infty} U_i = \{\mathbf{0}\}$ is not an open set. □

Solution to Exercise 10.9. *Show that no subset of $\mathbb{Q}^n$ containing more than one point is connected.*

Proof. Let $\mathbf{p} = (p_1, p_2, ..., p_n)$ and $\mathbf{q} == (q_1, q_2, ..., q_n)$ be distinct points in a set $S \subseteq \mathbb{Q}^n$, where $|\mathbf{p} - \mathbf{q}| = n\epsilon$. For each integer $1 \leq i \leq n$ choose irrational numbers s_i, t_i so that $p_i - \epsilon < s_i < p_i < t_i < p_i + \epsilon$. Let $U = \prod_{i=1}^{n}(s_i, t_i)$ and let $V = \mathbb{R}^{\ltimes} \setminus \overline{U}$. Then $U \cap S$ is non-empty since it contains $\mathbf{p}$, $V \cap S$ is non-empty since it contains $\mathbf{q}$. Both sets are open in S. Furthermore, if $\mathbf{z} = (z_1, z_2..., z_n)$ is a limit point of U which is not contained in U then some z_i must equal either s_i or t_i since if $z_i > t_i$ or $z_i < s_i$ then $\{(x, x_2..., x_n) \in \mathbb{R}^n | x_i > t_i$ or $x_i < s_i\}$ is an open set containing $\mathbf{z}$ which does not intersect U. But this means that $\mathbf{z} \notin S$ since one of its coordinates is not rational. Thus, U is both closed and open in S, and so is V, which means U and V are a separation of S. □

Solution to Exercise 10.10. *The empty set is the only proper subset of $\mathbb{R}^n$ which is both closed and open.*

Proof. First, we show that $\mathbb{R}^n$ is connected. Suppose $\mathbb{R}^n$ has a separation H, K and let $\mathbf{p} \in H$ and $\mathbf{q} \in K$. For some open sets U, V it follows that $H = U \cap \mathbb{R}^n = U$ and $K = U \cap \mathbb{R}^n = V$ which means that H and K are open sets. Define $f : [0, 1] \to \mathbb{R}^n$ by $f(x) = \mathbf{p} + t(\mathbf{q} - \mathbf{p})$. Then f is continuous, so since the continuous image a of a connected set is connected, $L(\mathbf{p}, \mathbf{q}) = f([0, 1])$ is connected. But then $H \cap L(\mathbf{p}, \mathbf{q})$ and $K \cap L(\mathbf{p}, \mathbf{q})$ are non-empty, their union is $L(\mathbf{p}, \mathbf{q})$ and the are open in $L(\mathbf{p}, \mathbf{q})$, making them a separation of $L(\mathbf{p}, \mathbf{q})$, which is a contradiction. We conclude that $\mathbb{R}^n$ is connected.

Let U be an open proper subset of $\mathbb{R}^n$. If U is also closed then $\mathbb{R}^n \setminus U$ is open, making $U, \mathbb{R}^n \setminus U$ a separation of $\mathbb{R}^n$, which is impossible since $\mathbb{R}^n$ is connected. □

Solution to Exercise 10.11. *Let $E \subseteq \mathbb{R}^n$. Then $\overline{E} = E \cup \partial(E)$.*

Proof. We know that $\overline{E}$ consists of the points of E and the limit points of E. Let $\mathbf{p}$ be a point which is not in E. Then every open set containing $\mathbf{p}$ contains a point not in E (namely $\mathbf{p}$). Thus, $\mathbf{p}$ is a limit point of E if and only if every open set containing $\mathbf{p}$ contains a point of E (since such a point will always be distinct from $\mathbf{p}$), and $\mathbf{p}$ is a boundary point of E if and only if every open set containing p contains a point of E. Since any point not in E is in the boundary of E if and only if it is in the closure of E, we see that $\overline{E} = E \cup \partial(E)$. □

Solution to Exercise 10.12. *If $A \subseteq K \subseteq \mathbb{R}^n$ and K is closed then A is closed in K if and only if A is closed.*

Proof. Let A be closed. Then $A \cap K = A$ is closed in K by definition. Assume A is closed in K. Then for some closed set B we know $B \cap K = A$. Since K is closed and the intersection of closed sets is closed, $B \cap K = A$ is a closed set. □

Solution to Exercise 10.13. *A point p of a metric space (X, d) is a limit point of a set $E \subseteq X$ if and only if every open set containing p contains infinitely many points of E.*

Proof. Suppose an open set U containing p contains at most finitely many points $q_1, q_2, ..., q_m$ of E distinct from p. Choose $\epsilon > 0$ so that $B_\epsilon(p) \subseteq U$. Let $\delta = \min\{\epsilon, d(p, q_1), d(p, q_2), d(p, q_3,), ..., d(p, q_m)\}$. Then $B_\delta(p)$ contains no points of E distinct from p, contradicting the assumption that p is a limit point of E. □

Chapter 11

Differentiation in Higher Dimensions

Definition 78

Let $f : V \to \mathbb{R}^m$, where V is an open set in $\mathbb{R}^n$. Then the *partial derivative* of f at $\mathbf{x} \in V$ with respect to the variable x_k is $f_{x_k}(\mathbf{x}) = \frac{\partial f}{\partial x_k}(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_k) - f(\mathbf{x})}{h}$. Let $f =< f_1, f_2, ..., f_m >$, where each $f_i : V \to \mathbb{R}$. The partial derivative of $f_{i_{x_p}}$ with respect to x_j is denoted $f_{i_{x_p x_j}} = \frac{\partial^2 f_i}{\partial x_j \partial x_p}$ for any natural numbers $p, j \leq n$. In a similar manner, we can inductively define $f_{i_{x_{j_1} x_{j_2} \dots x_{j_k}}} = \frac{\partial^k f_i}{\partial x_{j_k} \partial x_{j_{k-1}} ... \partial x_{j_1}}$ to be the kth partial derivative of f_i which is the derivative of $f_{i_{x_{j_1} x_{j_2} \dots x_{j_{k-1}}}}$ with respect to x_{j_k}. We say that f is C^k if all kth partial derivatives of all components of f are continuous. We say that f is C^∞ if f is C^k for all natural numbers k.

We also define mixed partials on the function f itself in the preceding definition in the same way, defining $f_{x_{j_1} x_{j_2} \dots x_{j_k}} = \frac{\partial^k f}{\partial x_{j_k} \partial x_{j_{k-1}} ... \partial x_{j_1}}$.

Partial derivatives are different from total derivatives of multivariable functions. The derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$, denoted $Df(\mathbf{x})$ is also referred to as the differential $df_{\mathbf{x}}$ in some texts, and is a linear transformation, usually written as an m by n matrix.

Definition 79

Let $f : V \to \mathbb{R}^m$, where V is an open set in $\mathbb{R}^n$. The *derivative* of f at $\mathbf{x}$ is the unique transformation matrix $Df(\mathbf{x})$ (or $df_{\mathbf{x}}$) satisfying $\lim_{\mathbf{h} \to \mathbf{0}} \frac{|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}|}{|\mathbf{h}|} = 0$ if such a transformation $Df(\mathbf{x})\mathbf{h}$ exists. We say that f is *differentiable* at $\mathbf{p}$ if f has a derivative at $\mathbf{p}$ and that f is differentiable if f is differentiable at every point of the domain of f.

If $f : V \to \mathbb{R}$ and the partial derivatives of f exist at a point $\mathbf{x} \in V$, then the *gradient* of f at $\mathbf{x}$ is the vector $\nabla f(\mathbf{x}) =< f_{x_1}, f_{x_2}, ..., f_{x_n} >$. So, if $z = f(x, y)$ then $\nabla f(a, b) =< f_x(a, b), f_y(a, b) >$, whereas if $w = f(x, y, z)$ then $\nabla f(a, b, c) =< f_x(a, b, c), f_y(a, b, c), f_z(a, b, c) >$.

If A is an $m \times n$ matrix and $T(\mathbf{x}) = A\mathbf{x}$ is a linear transformation from $\mathbb{R}^n$ into $\mathbb{R}^m$ then we denote $|A| = |T|$, the operator norm of T.

We use the notation $\triangle_f(\mathbf{x}) = \det Df(\mathbf{x})$ if $f : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable. This is referred to as the *Jacobian* of f at $\mathbf{x}$.

It turns out that if the derivative exists then it is $[\dfrac{\partial f_i}{\partial x_j}]_{m\times n}$. If the first partial derivatives are continuous at $\mathbf{x}$ then the derivative always exists (these things are shown below). Note that we also sometimes refer to the derivative as the transformation itself rather than the matrix generating the transformation by matrix multiplication. It is more convenient for us to think of the derivative as the matrix most of the time, with the understanding that the matrix is used to give the corresponding transformation.

Theorem 11.1. *Let $f : U \to \mathbb{R}^m$, where U is open in $\mathbb{R}^n$, $\boldsymbol{p} \in U$ and $f(\boldsymbol{x}) =< f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_m(\boldsymbol{x}) >$, where each component function $f_i : U \to \mathbb{R}$. Then f is differentiable if and only if each function f_i is differentiable.*

Proof. Let A be any $m \times n$ matrix with rows $A_1, A_2, .., A_m$ (listed from the top row to the bottom row). Then by Theorem 10.17, $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - A\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$ if and only if $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f_i(\mathbf{p}+\mathbf{h}) - f_i(\mathbf{p}) - A_i \cdot \mathbf{h}}{|\mathbf{h}|} = 0$ for each $1 \le i \le n$, which means that there is a matrix $A = Df(\mathbf{p})$ so that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - A\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$ if and only if, for each $1 \le i \le m$, there is a row vector (a $1 \times n$ matrix) A_i so that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f_i(\mathbf{p}+\mathbf{h}) - f_i(\mathbf{p}) - A_i \cdot \mathbf{h}}{|\mathbf{h}|} = 0$. The result follows. □

Theorem 11.2. *Let $f : U \to \mathbb{R}^m$, where U is open in $\mathbb{R}^n$, $\boldsymbol{p} \in U$ and let A be an $m \times n$ matrix. Then f is differentiable at $\boldsymbol{p}$ with derivative $A = Df(\boldsymbol{p})$ if and only if there is a function $\epsilon(\boldsymbol{h}) : U \to \mathbb{R}^m$ so that $f(\boldsymbol{p}+\boldsymbol{h}) - f(\boldsymbol{p}) = A\boldsymbol{h} + \epsilon(\boldsymbol{h})$ on U, where $\lim\limits_{\mathbf{h}\to\boldsymbol{0}} \dfrac{\epsilon(\boldsymbol{h})}{|\boldsymbol{h}|} = \boldsymbol{0}$.*

Proof. First, let f be differentiable at $\mathbf{p}$ with $A = Df(\mathbf{p})$. Let $\epsilon(\mathbf{h}) = f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - Df(\mathbf{p})\mathbf{h}$. By definition of derivative, we know that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\epsilon(\mathbf{h})}{|\mathbf{h}|} = \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - Df(\mathbf{p})\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$.

Next, assume there is a function $\epsilon(\mathbf{h}) : U \to \mathbb{R}^m$ so that $f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) = A\mathbf{h} + \epsilon(\mathbf{h})$ on U for some matrix A, where $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\epsilon(\mathbf{h})}{|\mathbf{h}|} = \mathbf{0}$. Then that means $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\epsilon(\mathbf{h})}{|\mathbf{h}|} = \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - A\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$, so f is differentiable and $A = Df(\mathbf{p})$.

□

Theorem 11.3. *Let $f : V \to \mathbb{R}^m$, where V is an open subset of $\mathbb{R}^n$ and $\boldsymbol{p} \in V$ and f is differentiable at $\boldsymbol{p}$. Then f is continuous at $\boldsymbol{x}$.*

Proof. By Theorem 11.2, there is a function $\epsilon(\mathbf{h}) : V \to \mathbb{R}^m$ so that $f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) = Df(\mathbf{p})\mathbf{h} + \epsilon(\mathbf{h})$ on U, where $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\epsilon(\mathbf{h})}{|\mathbf{h}|} = \mathbf{0}$. Hence, $\lim\limits_{\mathbf{h}\to\mathbf{0}} f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) = \lim\limits_{\mathbf{h}\to\mathbf{0}} Df(\mathbf{p})\mathbf{h} + \epsilon(\mathbf{h}) = \mathbf{0}$, which means that $\lim\limits_{\mathbf{h}\to\mathbf{0}} f(\mathbf{p}+\mathbf{h}) = f(\mathbf{p})$ and thus $\lim\limits_{\mathbf{x}\to\mathbf{p}} f(\mathbf{x}) = f(\mathbf{p})$, so f is continuous at $\mathbf{p}$. □

Theorem 11.4. *Let $f : V \to \mathbb{R}$, where V is an open subset of $\mathbb{R}^n$ and $\boldsymbol{x} \in V$ and f is differentiable at $\boldsymbol{x}$. Then $Df(\boldsymbol{x}) = \triangledown f(\boldsymbol{x})$.*

Proof. We know that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\cdot\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$. Thus, in particular, for any given $1 \le i \le n$ we know that if $\mathbf{h} = \mathbf{e}_i$ then $\lim\limits_{h\to 0} \dfrac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x}) - Df(\mathbf{x})\cdot h\mathbf{e}_i}{h} = 0$. Hence, if $Df(\mathbf{x})_i$ is the ith coordinate of $Df(\mathbf{x})$ then $\lim\limits_{h\to 0} \dfrac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x}) - hDf(\mathbf{x})_i}{h} = 0$ and $\lim\limits_{h\to 0} \dfrac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h} = Df(\mathbf{x})_i$. Thus, by definition, $f_{x_i}(\mathbf{x})$ exists and is equal to $Df(\mathbf{x})_i$ for each $1 \le i \le n$, so $Df(\mathbf{x}) = \triangledown f(\mathbf{x})$.

□

Theorem 11.5. *Let $f : V \to \mathbb{R}^m$ be differentiable at $\boldsymbol{x}$, where V is an open subset of $\mathbb{R}^n$ and $\boldsymbol{x} \in V$. Then $Df(\boldsymbol{x}) = [\dfrac{\partial f_i}{\partial x_j}]_{m\times n}$.*

Proof. We know that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}|}{|\mathbf{h}|} = 0$, which is true if and only if $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x})\cdot\mathbf{h}}{|\mathbf{h}|} = 0$ for each i, where $Df_i(\mathbf{x})$ represents the ith row of $Df(\mathbf{x})$. By theorem 11.4, we know that $Df_i(\mathbf{x}) = \triangledown f_i(\mathbf{x})$, which means that $Df(\mathbf{x}) = [\dfrac{\partial f_i}{\partial x_j}]_{m\times n}$. □

Theorem 11.6. *Let $f : V \to \mathbb{R}$, where V is an open subset of $\mathbb{R}^n$ and $\boldsymbol{x} \in V$ so that the partial derivatives of f exist on V and are continuous at $\boldsymbol{x}$.*

(a) Then f is differentiable at $\boldsymbol{x}$.

(b) If the partials of f are continuous on V and K is a compact subset of V then for every $\epsilon > 0$ there is a $\delta > 0$ so that if $\boldsymbol{x} \in K$ and $|\boldsymbol{h}| < \delta$ then $f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \triangledown f(\boldsymbol{x})\cdot\boldsymbol{h} + R(\boldsymbol{h})$, where $|R(\boldsymbol{h})| < \epsilon|\boldsymbol{h}|$.

Proof. We know that if the derivative exists then it equals the gradient. We will show that the gradient satisfies the definition of derivative. Choose $r > 0$ so that $B_r(\mathbf{x}) \subset V$ and let $\mathbf{h}$ be a vector so that $|\mathbf{h}| < r$, where $\mathbf{h} =< h_1, h_2, ..., h_n >$ and $\mathbf{x} =< x_1, x_2, ..., x_n >$. Note that

$f(\mathbf{x}+\mathbf{h})-f(\mathbf{x}) = (f(x_1+h_1, x_2+h_2, ..., x_n+h_n)-f(x_1+h_1, x_2+h_2, ..., x_{n-1}+h_{n-1}, x_n)) + (f(x_1+h_1, x_2+h_2, ..., x_{n-1}+h_{n-1}, x_n)-f(x_1+h_1, x_2+h_2, ..., x_{n-2}+h_{n-2}, x_{n-1}, x_n)) + ... + (f(x_1+h_1, x_2, ..., x_n)-f(x_1, x_2, ..., x_n))$. For each $1 \leq i \leq n$ we define $g_i(t) = f(x_1+h_1, x_2+h_2, ..., x_i+t, x_{i+1}, ..., x_n)$. Note that $g'(t) = \lim_{h\to 0} \dfrac{f(x_1+h_1, ..., x_i+t+h, x_{i+1}, ..., x_n)-f(x_1+h_1, ..., x_i+t, x_{i+1}, ..., x_n)}{h} = f_{x_i}(x_1+h_1, ..., x_i+t, x_{i+1}, ..., x_n)$. Since g is a function of one variable, by the Mean Value Theorem we can find, for each such integer i, a point c_i between 0 and h_i, so that $g'(c_i)(h_i-0) = g(h_i)-g(0) = f(x_1+h_1, x_2+h_2, ..., x_i+h_i, x_{i+1}, ..., x_n)- f(x_1+h_1, x_2+h_2, ..., x_i, x_{i+1}, ..., x_n)$ unless $h_i = 0$ in which case the Mean Value Theorem does not apply, and we set $c_i = x_i$ and the equation $g'(c_i)(h_i-0) = g(h_i)-g(0) = f(x_1+h_1, x_2+h_2, ..., x_i+h_i, x_{i+1}, ..., x_n)- f(x_1+h_1, x_2+h_2, ..., x_i, x_{i+1}, ..., x_n)$ reduces to $0 = 0 = 0$, which is still true.

Thus, it follows that $f(\mathbf{x}+\mathbf{h})-f(\mathbf{x}) = \sum_{i=1}^{n} g'(c_i)h_i = \sum_{i=1}^{n} f_{x_i}(x_1+h_1, x_2+h_2, ..., x_i+c_i, x_{i+1}, ..., x_n)h_i = \delta(\mathbf{h})\cdot\mathbf{h}$, where $\delta(\mathbf{h})$ is the vector whose ith coordinate is $f_{x_i}(x_1+h_1, x_2+h_2, ..., x_i+c_i, x_{i+1}, ..., x_n)$. Thus, $f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h} = \sum_{i=1}^{n}(f_{x_i}(x_1+h_1, x_2+h_2, ..., x_i+c_i, x_{i+1}, ..., x_n)-f_{x_i}(x_1, x_2, ..., x_n))h_i = \mathbf{h}\cdot(\delta(\mathbf{h})-\nabla f(\mathbf{x}))$. But since the c_i are between 0 and h_i we know that each c_i approaches zero and each corresponding x_i+c_i approaches x_i as $\mathbf{h}$ approaches zero. Since each f_{x_i} is continuous at $\mathbf{x}$, it follows that $\lim_{\mathbf{h}\to\mathbf{0}} \delta(\mathbf{h})-\nabla f(\mathbf{x}) = \mathbf{0}$. By the Cauchy Schwarz inequality we know that $\dfrac{|\mathbf{h}\cdot(\delta(\mathbf{h})-\nabla f(\mathbf{x}))|}{|\mathbf{h}|} \leq \dfrac{|\mathbf{h}||\delta(\mathbf{h})-\nabla f(\mathbf{x})|}{|\mathbf{h}|} = |\delta(\mathbf{h})-\nabla f(\mathbf{x})|$, so $\lim_{\mathbf{h}\to\mathbf{0}} \dfrac{|f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h}|}{|\mathbf{h}|} = 0$ by the Squeeze Theorem.

(b) For each $\mathbf{x} \in K$ choose an $\epsilon_{\mathbf{x}} > 0$ so that $\overline{B_{\epsilon_{\mathbf{x}}}(\mathbf{x})} \subset V$. Since K is compact and $\{B_{\epsilon_{\mathbf{x}}}(\mathbf{x})\}_{\mathbf{x}\in K}$ is an open cover of K, we can find a finite subcover $F = \{B_{\epsilon_{\mathbf{x}_i}}(\mathbf{x}_i)\}_{1\leq i\leq k}$.

By the Lebesgue Number Lemma we can find a $\gamma > 0$ so that if S is a set in $\mathbb{R}^n$ whose diameter does not exceed γ and $S \cap K \neq \emptyset$ then $S \subset B_{\epsilon_{\mathbf{x}_i}}(\mathbf{x}_i)$ for some i. Note also that $H = \bigcup_{i=1}^{k} \overline{B_{\epsilon_{\mathbf{x}_i}}(\mathbf{x}_i)}$ is a compact subset of V.

For each $i \in \{1, 2, ..., n\}$, the partials f_{x_i} are uniformly continuous on H by Theorem 10.35, so we can choose $\delta_i > 0$ so that if $|\mathbf{x}-\mathbf{y}| < \delta_i$ and $\mathbf{x}, \mathbf{y} \in H$ then $|f_{x_i}(\mathbf{x})-f_{x_i}(\mathbf{y})| < \dfrac{\epsilon}{\sqrt{n}}$. Let $\delta = \min\{\gamma, \delta_1, \delta_2, ..., \delta_n\}$.

We note then, from the proof of (a), that $|\delta(\mathbf{h})_i - f_{x_i}(\mathbf{x})| < \dfrac{\epsilon}{\sqrt{n}}$ for each $i \in \{1, 2, ..., n\}$ if $|\mathbf{h}| < \delta$. As in part (a) we see that $\dfrac{|f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h}|}{|\mathbf{h}|} = \dfrac{|\mathbf{h}\cdot(\delta(\mathbf{h})-\nabla f(\mathbf{x}))|}{|\mathbf{h}|} \leq \dfrac{|\mathbf{h}||\delta(\mathbf{h})-\nabla f(\mathbf{x})|}{|\mathbf{h}|} = |\delta(\mathbf{h})-\nabla f(\mathbf{x})| = \sqrt{\sum_{i=1}^{n}(\delta(\mathbf{h})_i - f_{x_i}(\mathbf{x}))^2} < \sqrt{n}\dfrac{\epsilon}{\sqrt{n}} = \epsilon$.

Since $\dfrac{|f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h}|}{|\mathbf{h}|} < \epsilon$, if we set $R(\mathbf{h}) = f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h}$ then we see that $|R(\mathbf{h})| = |f(\mathbf{x}+\mathbf{h})-f(\mathbf{x})-\nabla f(\mathbf{x})\cdot\mathbf{h}| < \epsilon|\mathbf{h}|$, and $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})\cdot\mathbf{h} + R(\mathbf{h})$.

□

We next extend the previous result to functions from open sets in $\mathbb{R}^n$ into $\mathbb{R}^m$ (instead of just $\mathbb{R}$).

Theorem 11.7. *Let $f : V \to \mathbb{R}^m$, where V is an open subset of $\mathbb{R}^n$ and $\boldsymbol{x} \in V$ so that the partial derivatives of f exist on V and are continuous at $\boldsymbol{x}$. Then:*

(a) f is differentiable at $\boldsymbol{x}$.

(b) If the partials of f are continuous on V and K is a compact subset of V then for every $\epsilon > 0$ there is a $\delta > 0$ so that if $\boldsymbol{x} \in K$ and $|\boldsymbol{h}| < \delta$ then $f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + Df(\boldsymbol{x})\boldsymbol{h} + R(\boldsymbol{h})$, where $|R(\boldsymbol{h})| < \epsilon|\boldsymbol{h}|$.

Proof. (a) By theorem 11.6, we know that for each $1 \leq i \leq m$ it is the case that f_i is differentiable at $\mathbf{x}$, which means that f is differentiable at $\mathbf{x}$ by Theorem 11.1.

(b) By 11.6 we know that for each $i \in \{1, 2, 3, ..., n\}$ we can find $\delta_i > 0$ so that if $\mathbf{x} \in K$ and $|\mathbf{h}| < \delta_i$ then $f_i(\mathbf{x} + \mathbf{h}) = f_i(\mathbf{x}) + \triangledown f_i(\mathbf{x}) \cdot \mathbf{h} + R_i(\mathbf{h})$, where $|R_i(\mathbf{h})| < \frac{\epsilon}{\sqrt{n}}|\mathbf{h}|$, so that $\frac{|f_i(\mathbf{x} + \mathbf{h}) - f_i(\mathbf{x}) - \triangledown f_i(\mathbf{x}) \cdot \mathbf{h}|}{|\mathbf{h}|} < \frac{\epsilon}{\sqrt{n}}$ for each $i \in \{1, 2, 3, ..., n\}$.

Let $\delta = \min\{\delta_1, \delta_2, ..., \delta_n\}$. If $|\mathbf{h}| < \delta$ it follows that $\frac{|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}|}{|\mathbf{h}|} = \sqrt{\sum_{i=1}^{n} \frac{(f_i(\mathbf{x} + \mathbf{h}) - f_i(\mathbf{x}) - \triangledown f_i(\mathbf{x}) \cdot \mathbf{h})^2}{|\mathbf{h}|^2}} < \epsilon$.

Set $R(\mathbf{h}) = f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}$. Then $|R(\mathbf{h})| = |f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}| < \epsilon|\mathbf{h}|$, and $f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + R(\mathbf{h})$. □

Theorem 11.8. *Clairaut's Theorem. Let $f : V \to \mathbb{R}$ be C^2, where V is an open set in $\mathbb{R}^2$. Then $f_{xy} = f_{yx}$ on V.*

Proof. Let $\mathbf{a} = (x_0, y_0) \in V$. Since V is open we can find an $\epsilon > 0$ so that $B_\epsilon(\mathbf{a}) \subset V$. Then for h, k so that $|h|, |k| < \frac{\epsilon}{2}$ it follows that $(x_0 + h, y_0 + k) \in V$ and we can define $d(h, k) = f(x_0 + h, y_0 + k) - f(x_0, y_0 + k) - f(x_0 + h, y_0) + f(x_0, y_0)$. Define $g_1(t) = f(x_0 + th, y_0 + k) - f(x_0 + th, y_0)$. Then g_1 is differentiable and $g_1(1) - g_1(0) = d(h, k)$ and by the Mean Value Theorem there is a point $c_h \in (0, 1)$ so that $g_1'(c_h)(1 - 0) = g_1(1) - g_1(0) = d(h, k)$. But by the chain rule, $g_1'(c_h) = h(f_x(x_0 + c_h h, y_0 + k) - f_x(x_0 + c_h h, y_0))$, so $\frac{d(h, k)}{h} = f_x(x_0 + c_h h, y_0 + k) - f_x(x_0 + c_h h, y_0)$. Next define $g_2(t) = f_x(x_0 + c_h h, y_0 + tk)$ and note that $g_2(1) - g_2(0) = f_x(x_0 + c_h h, y_0 + k) - f_x(x_0 + c_h h, y_0)$, so by the Mean Value Theorem there is a point $c_k \in (0, 1)$ so that $g_2'(c_k)(1 - 0) = g_2(1) - g_2(0) = \frac{d(h, k)}{h}$. By the chain rule, we know that $g_2'(c_k) = kf_{xy}(x_0 + c_h h, y_0 + c_k k)$, which means that $f_{xy}(x_0 + c_h h, y_0 + c_k k) = \frac{d(h, k)}{hk}$. Since f is C^2 on V we know that $\lim_{(h,k)\to(0,0)} f_{xy}(x_0 + c_h h, y_0 + c_k k) = f_{xy}(x_0, y_0) = \lim_{(h,k)\to(0,0)} \frac{d(h, k)}{hk}$.

We then define $g_3(t) = f(x_0 + h, y_0 + tk) - f(x_0, y_0 + tk)$ and note that $g_3(1) - g_3(0) = d(h, k)$ and by the Mean Value Theorem there is a point $q_k \in (0, 1)$ so that $g_3'(q_k)(1 - 0) = g_3(1) - g_3(0) = d(h, k)$. But by the chain rule, $g_3'(q_k) = k(f_y(x_0 + h, y_0 + q_k k) - f_y(x_0, y_0 +$

$q_k k)$), so $\dfrac{d(h,k)}{k} = f_y(x_0+h, y_0+q_k k) - f_y(x_0, y_0+q_k k)$. Finally, let $g_4(t) = f_y(x_0+th, y_0+q_k k)$ and note that $g_4(1) - g_4(0) = f_y(x_0+h, y_0+q_k k) - f_y(x_0, y_0+q_k k)$, so by the Mean Value Theorem there is a point $q_h \in (0,1)$ so that $g_4'(q_h)(1-0) = g_4(1) - g_4(0) = \dfrac{d(h,k)}{k}$. By the chain rule, we know that $g_4'(q_h) = hf_{yx}(x_0+q_h h, y_0+q_k k)$, which means that $f_{yx}(x_0+q_h h, y_0+q_k k) = \dfrac{d(h,k)}{hk}$. Since f is C^2 on V we know that $\lim\limits_{(h,k)\to(0,0)} f_{yx}(x_0 + c_h h, y_0+c_k k) = f_{yx}(x_0,y_0) = \lim\limits_{(h,k)\to(0,0)} \dfrac{d(h,k)}{hk} = f_{xy}(x_0,y_0)$.

□

Since partial derivatives are calculated with other variables fixed, we can also extend Clairaut's Theorem to switching the order of any pair of variables for a multivariable mixed partial. This is addressed in the exercises.

It is sometimes useful to use the following alternate definition of differentiability, which we show is equivalent to the one we gave. Either can be considered the definition of differentiable for a function from $\mathbb{R}^n$ into $\mathbb{R}$.

Theorem 11.9. *Let $f : V \to \mathbb{R}$ where V is open in $\mathbb{R}^n$. Then f is differentiable at $\boldsymbol{x} \in V$ if and only if there are functions $\epsilon_i(\boldsymbol{h})$ for $1 \le i \le n$ so that $f(\boldsymbol{x}+\boldsymbol{h}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}) \cdot \boldsymbol{h} = \sum\limits_{i=1}^{n} \epsilon_i(\boldsymbol{h}) h_i$ where $\lim\limits_{\boldsymbol{h}\to\boldsymbol{0}} \epsilon_i(\boldsymbol{h}) = 0$ for each i.*

Proof. First, we assume that f is differentiable at $\mathbf{x}$. Then $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}}{|\mathbf{h}|} = 0$. We then define
$\epsilon_i(\mathbf{h}) = \dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}}{\sum_{i=1}^{n} |h_i|}$ if $h_i \ge 0$ and
$\epsilon_i(\mathbf{h}) = -\dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}}{\sum_{i=1}^{n} |h_i|}$ if $h_i < 0$. Then
$\sum\limits_{i=1}^{n} \epsilon_i(\mathbf{h}) h_i = \dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}}{\sum_{i=1}^{n} |h_i|} \sum\limits_{i=1}^{n} |h_i| = f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}$. Since $\sum\limits_{i=1}^{n} |h_i| \ge |\mathbf{h}|$ and $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}|}{|\mathbf{h}|} = 0$, it follows that
$\lim\limits_{\mathbf{h}\to\mathbf{0}} |\epsilon_i(\mathbf{h})| = \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{|f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}|}{\sum_{i=1}^{n} |h_i|} = 0$, so $\lim\limits_{\mathbf{h}\to\mathbf{0}} \epsilon_i(\mathbf{h}) = 0$.

Next, assume that there are functions $\epsilon_i(\mathbf{h})$ for $1 \le i \le n$ so that $f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h} = \sum\limits_{i=1}^{n} \epsilon_i(\mathbf{h}) h_i$ where $\lim\limits_{\mathbf{h}\to\mathbf{0}} \epsilon_i(\mathbf{h}) = 0$ for each i.

Then $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \mathbf{h}}{|\mathbf{h}|} = \lim\limits_{\mathbf{h}\to\mathbf{0}} \sum\limits_{i=1}^{n} \dfrac{\epsilon_i(\mathbf{h}) h_i}{|\mathbf{h}|} = 0$ since $\dfrac{|h_i|}{|\mathbf{h}|} \le 1$ and $\lim\limits_{\mathbf{h}\to\mathbf{0}} \epsilon_i(\mathbf{h}) = 0$ for each i. Thus, f is differentiable.

□

Sometimes it looks nicer to represent the coordinates of $\mathbf{h}$ in the preceding theorem as changes in the variables. So, in two coordinates you might represent this vector as $\mathbf{h} = (\triangle x, \triangle y)$ for instance. Using this format, for the specific case of $n = 2$ and $n = 3$ theorem 11.9 can be stated as follows.

Let $f : V \to \mathbb{R}$ where V is open in $\mathbb{R}^2$. Then f is differentiable at $(x, y) \in V$ if and only if there are functions $\epsilon_1(\triangle x, \triangle y)$ and $\epsilon_2(\triangle x, \triangle y)$ so that $f(x+\triangle x, y+\triangle y) - f(x, y) - f_x(x, y)\triangle x - f_y(x, y)\triangle y = \epsilon_1(\triangle x, \triangle y)\triangle x + \epsilon_2(\triangle x, \triangle y)\triangle y$ where $\lim\limits_{(\triangle x,\triangle y)\to(0,0)} \epsilon_1(\triangle x, \triangle y) = 0 = \lim\limits_{(\triangle x,\triangle y)\to(0,0)} \epsilon_2(\triangle x, \triangle y)$.

Let $f : V \to \mathbb{R}$ where V is open in $\mathbb{R}^3$. Then f is differentiable at $(x, y, z) \in V$ if and only if there are functions $\epsilon_1(\triangle x, \triangle y, \triangle z)$, $\epsilon_2(\triangle x, \triangle y, \triangle z)$ and $\epsilon_3(\triangle x, \triangle y, \triangle z)$ so that $f(x + \triangle x, y + \triangle y, z + \triangle z) - f(x, y, z) - f_x(x, y, z)\triangle x - f_y(x, y, z)\triangle y - f_z(x, y, z)\triangle z = \epsilon_1(\triangle x, \triangle y, \triangle z)\triangle x + \epsilon_2(\triangle x, \triangle y, \triangle z)\triangle y + \epsilon_3(\triangle x, \triangle y, \triangle z)\triangle z$ where $\lim\limits_{(\triangle x,\triangle y,\triangle z)\to(0,0,0)} \epsilon_1 = \lim\limits_{(\triangle x,\triangle y,\triangle z)\to(0,0,0)} \epsilon_2 = \lim\limits_{(\triangle x,\triangle y,\triangle z)\to(0,0,0)} \epsilon_3 = 0$.

Theorem 11.10. *Chain Rule for Euclidean Spaces. Let $f : U \to \mathbb{R}^m$ be differentiable at $\boldsymbol{p}$ and let $g : V \to \mathbb{R}^j$ be differentiable at $g(\boldsymbol{p})$, where U is open in $\mathbb{R}^n$ and V is open in $\mathbb{R}^m$. Then $g \circ f$ is differentiable at $\boldsymbol{p}$ and $D(g \circ f)(\boldsymbol{p}) = Dg(f(\boldsymbol{p}))Df(\boldsymbol{p})$.*

Proof. First, there is an $R > 0$ so that $B_R(f(\mathbf{p})) \subset V$, and we can choose $r > 0$ so that if $|\mathbf{p} - \mathbf{x}| < r$ then $|f(\mathbf{x}) - f(\mathbf{p})| < R$ (since f is continuous at $\mathbf{p}$), and $B_r(\mathbf{p}) \subset U$, which means that $B_r(\mathbf{p})$ is contained in the domain of $g \circ f$. For the remainder of the argument, assume that $0 < |\mathbf{h}| < r$ and $\mathbf{h} \in \mathbb{R}^n$.

We also observe that $Dg(f(\mathbf{p}))Df(\mathbf{p})$ is a $j \times n$ matrix, which is the correct matrix size for the derivative.

By Theorem 11.2, we can find $\epsilon(\mathbf{h}) : U \to \mathbb{R}^m$ so that $f(\mathbf{p}+\mathbf{h})-f(\mathbf{p}) = Df(\mathbf{p})\mathbf{h}+\epsilon(\mathbf{h})$ on U, where $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{\epsilon(\mathbf{h})}{|\mathbf{h}|} = \mathbf{0}$. Likewise, we can find $\delta(\mathbf{k}) : V \to \mathbb{R}^j$ so that $g(f(\mathbf{p})+\mathbf{k})-g(f(\mathbf{p})) = Dg(f(\mathbf{p}))\mathbf{k} + \delta(\mathbf{k})$ and $\lim\limits_{\mathbf{k}\to\mathbf{0}} \dfrac{\delta(\mathbf{k})}{|\mathbf{k}|} = \mathbf{0}$.

Note $g(f(\mathbf{p} + \mathbf{h})) = g(f(\mathbf{p}) + (f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}))$.

Thus, using $\mathbf{k} = f(\mathbf{p}+\mathbf{h})-f(\mathbf{p})$, we can rewrite $g(f(\mathbf{p}+\mathbf{h}))-g(f(\mathbf{p})) = Dg(f(\mathbf{p}))(f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p})) + \delta(f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}))$.

Hence, replacing the $f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p})$ with $Df(\mathbf{p})\mathbf{h} + \epsilon(\mathbf{h})$, see that
$g(f(\mathbf{p} + \mathbf{h})) - g(f(\mathbf{p})) - Dg(f(\mathbf{p}))Df(\mathbf{p})\mathbf{h}$ can be written as
$Dg(f(\mathbf{p}))(Df(\mathbf{p})\mathbf{h} + \epsilon(\mathbf{h})) + \delta(f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p})) - Dg(f(\mathbf{p}))Df(\mathbf{p})\mathbf{h} =$
$Dg(f(\mathbf{p}))(\epsilon(\mathbf{h})) + \delta(f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}))$.

Thus, $\dfrac{|g(f(\mathbf{p} + \mathbf{h})) - g(f(\mathbf{p})) - Dg(f(\mathbf{p}))Df(\mathbf{p})\mathbf{h}|}{|\mathbf{h}|} \leq$
$\dfrac{|Dg(f(\mathbf{p}))(\epsilon(\mathbf{h}))|}{|\mathbf{h}|} + \dfrac{|\delta(f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}))|}{|\mathbf{h}|}$. By Theorem 10.32, we know that
$\dfrac{|Dg(f(\mathbf{p}))(\epsilon(\mathbf{h}))|}{|\mathbf{h}|} \leq \dfrac{|Dg(f(\mathbf{p}))||(\epsilon(\mathbf{h}))|}{|\mathbf{h}|}$.

Let $\gamma > 0$. Since $\delta(\mathbf{0}) = \mathbf{0}$, it is true that when $f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) = \mathbf{0}$, we know that $\dfrac{|\delta(f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}))|}{|\mathbf{h}|} = 0 < \dfrac{\gamma}{2}$. If $f(\mathbf{p} + \mathbf{h}) - f(\mathbf{p}) \neq \mathbf{0}$ then

$$\frac{|\delta(f(\mathbf{p}+\mathbf{h})-f(\mathbf{p}))|}{|f(\mathbf{p}+\mathbf{h})-f(\mathbf{p}))|}\frac{|f(\mathbf{p}+\mathbf{h})-f(\mathbf{p})|}{|\mathbf{h}|}=\frac{|\delta(f(\mathbf{p}+\mathbf{h})-f(\mathbf{p}))|}{|\mathbf{h}|}.$$

Since we know that $\lim\limits_{\mathbf{h}\to 0}\frac{|(\epsilon(\mathbf{h}))|}{|\mathbf{h}|}=0$, we can find $t\in(0,r)$ so that if $0<|\mathbf{h}|<t$ then $\frac{|Dg(f(\mathbf{p}))||(\epsilon(\mathbf{h}))|}{|\mathbf{h}|}<\frac{\gamma}{2}$.

Next observe that $|f(\mathbf{p}+\mathbf{h})-f(\mathbf{p})|=|Df(\mathbf{p})\mathbf{h}+\epsilon(\mathbf{h})|\leq|Df(\mathbf{p})||\mathbf{h}|+|\epsilon(\mathbf{h})|$. Since $\lim\limits_{\mathbf{h}\to\mathbf{0}}\frac{|\epsilon(\mathbf{h})|}{|\mathbf{h}|}=0$ we can find $s\in(0,t)$ so that if $|\mathbf{h}|<s$ then $\frac{|\epsilon(\mathbf{h})|}{|\mathbf{h}|}<1$, which means that $\frac{|f(\mathbf{p}+\mathbf{h})-f(\mathbf{p})|}{|\mathbf{h}|}<|Df(\mathbf{p})|+1$.

Choose $0<\beta<R$ so that if $0<|\mathbf{k}|<\beta$ then $\frac{|\delta(\mathbf{k})|}{|\mathbf{k}|}<\frac{\gamma}{2(|Df(\mathbf{p})|+1)}$. Choose $0<\alpha<s$ so that if $0<|\mathbf{h}|<\alpha$ then $|f(\mathbf{p}+\mathbf{h})-f(\mathbf{p})|<\beta$, so it follows that $\frac{|g(f(\mathbf{p}+\mathbf{h}))-g(f(\mathbf{p}))-Dg(f(\mathbf{p}))Df(\mathbf{p})\mathbf{h}|}{|\mathbf{h}|}\leq\frac{|Dg(f(\mathbf{p}))(\epsilon(\mathbf{h}))|}{|\mathbf{h}|}+\frac{|\delta(f(\mathbf{p}+\mathbf{h})-f(\mathbf{p}))|}{|\mathbf{h}|}<\frac{\gamma}{2}+\frac{\gamma}{2(|Df(\mathbf{p})|+1)}(|Df(\mathbf{p})|+1)=\gamma$. We conclude that $\lim\limits_{\mathbf{h}\to\mathbf{0}}\frac{|g(f(\mathbf{p}+\mathbf{h}))-g(f(\mathbf{p}))-Dg(f(\mathbf{p}))Df(\mathbf{p})\mathbf{h}|}{|\mathbf{h}|}=0$, so $D(g\circ f)(\mathbf{p})=Dg(f(\mathbf{p}))Df(\mathbf{p})$ as desired.

□

One helpful consequence of the Chain Rule is a way of simplifying implicit differentiation.

Definition 80

If $\mathbf{x}\in\mathbb{R}^s$ and $\mathbf{y}\in\mathbb{R}^t$ we use the notation $(\mathbf{x},\mathbf{y})=(x_1,x_2,...,x_s,y_1,y_2,...,y_t)\in\mathbb{R}^{s+t}$. The *graph* of $F(\mathbf{x})=k$ is the set of solutions of the equation $F(\mathbf{x})=k$. The graph of a function $z=f(\mathbf{x})$ over a domain $D\subseteq\mathbb{R}^n$ is $\{(\mathbf{x},z)\in\mathbb{R}^{n+1}|z=f(\mathbf{x})$ and $\mathbf{x}\in D\}$. We say a set $S\subseteq\mathbb{R}^{n+1}$ is *locally the graph of a function near (or at)* $\boldsymbol{x}_0$ if there is an $\epsilon>0$ so that $B_\epsilon(\mathbf{p})\cap S$ is the graph of a function of n variables, meaning that it is the graph of a function $x_j=f(x_1,x_2,...,x_{j-1},x_{j+1},...,x_{n+1})$ for some $1\leq j\leq n+1$, in which case we say that S (or an equation whose solutions are S) defined x_j as a function of $x_1,x_2,...,x_{j-1},x_{j+1},...,x_{n+1}$ near (or at) $\mathbf{x}_0=(x_1,x_2,...,x_{n+1})$ or on the ball of radius ϵ about $\mathbf{x}_0$. We would say that S is locally the graph of a function if it is locally the graph of a function at every point of S. Similarly, we would say that S is locally the graph of a differentiable or C^k function at $\mathbf{x}_0$ of the function f is differentiable or C^k.

We say a a vector $\mathbf{n}\in\mathbb{R}^{n+1}$ is *normal* to S (or orthogonal to S or perpendicular to S) at a point $\mathbf{p}\in S$ which is a limit point of S if for every sequence $\{\mathbf{x}_i\}\subseteq S$ so that $\{\mathbf{x}_i\}\to\mathbf{p}$ it is the case that $\{\frac{\mathbf{x}_i-\mathbf{p}}{|\mathbf{x}_i-\mathbf{p}|}\cdot\mathbf{n}\}\to 0$. If $\mathbf{n}$ is normal to S at $\mathbf{p}$ then we say that $P=\{\mathbf{x}\in R^{n+1}|\mathbf{n}\cdot(\mathbf{x}-\mathbf{p})=0\}$ is the *tangent hyperplane* to S at $\mathbf{p}$. If $n=2$ then we refer to P as the tangent plane to S at $\mathbf{p}$.

Theorem 11.11. *Let S be the graph of $z = f(x_1, x_2, ..., x_n)$ for a function $f : D \to \mathbb{R}$ on some domain $D \subseteq \mathbb{R}^n$ containing the ball $B_\epsilon(\boldsymbol{c})$ so that f is differentiable at $\boldsymbol{c}$. Then the vector $\boldsymbol{n} =< f_{x_1}(\boldsymbol{c}), f_{x_2}(\boldsymbol{c}), ..., f_{x_n}(\boldsymbol{c}), -1 >$ is normal to S at $(\boldsymbol{c}, f(\boldsymbol{c})) = \boldsymbol{p}$ and has corresponding tangent hyperplane $P = \{\boldsymbol{x} \in R^{n+1} | \boldsymbol{n} \cdot (\boldsymbol{x} - \boldsymbol{p}) = 0\}$.*

Proof. Let $\{\mathbf{x}_i\}$ be a sequence of points in $S \setminus \mathbf{p}$ which converges to $\mathbf{p} = (\mathbf{c}, f(\mathbf{c}))$. Then for each natural number i we can find $\mathbf{h}_i$ so that $\mathbf{x}_i = (\mathbf{c} + \mathbf{h}_i, f(\mathbf{c} + \mathbf{h}_i))$. Since f is differentiable at $\mathbf{c}$ we know that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{c}+\mathbf{h}) - f(\mathbf{c}) - \nabla f(\mathbf{c}) \cdot \mathbf{h}}{|\mathbf{h}|} = 0$, so by the sequential characterization of limits we know that $\{\dfrac{f(\mathbf{c}+\mathbf{h}_i) - f(\mathbf{c}) - \nabla f(\mathbf{c}) \cdot \mathbf{h}_i}{|\mathbf{h}_i|}\} \to 0$. Since $\mathbf{x}_i = (\mathbf{c} + \mathbf{h}_i, f(\mathbf{c} + \mathbf{h}_i))$ it follows that $\mathbf{x}_i - \mathbf{p} = (\mathbf{h}_i, f(\mathbf{c} + \mathbf{h}_i) - f(\mathbf{c}))$, so $\{\dfrac{\mathbf{x}_i - \mathbf{p}}{|\mathbf{x}_i - \mathbf{p}|} \cdot \mathbf{n}\} = \{\dfrac{\nabla f(\mathbf{c}+\mathbf{h}_i) \cdot \mathbf{h}_i - (f(\mathbf{c}+\mathbf{h}_i) - f(\mathbf{c}))}{|\mathbf{h}_i|}\} \to 0$. The result follows. □

Theorem 11.12. *Let S be the graph of $F(\boldsymbol{x}) = k$, where S is locally the graph of a differentiable function $z = f(\boldsymbol{x})$ at $\boldsymbol{p} = (\boldsymbol{x}_0, f(\boldsymbol{x}_0)) \in \mathbb{R}^{m+1}$, where $\boldsymbol{x}_0 = (x_1, x_2, x_3, ..., x_m) \in \mathbb{R}^m$. Then $\dfrac{\partial x_i}{\partial x_j}(\boldsymbol{x}_0) = -\dfrac{F_{x_j}(\boldsymbol{x}_0)}{F_{x_i}(\boldsymbol{x}_0)}$ if $F_{x_i}(\boldsymbol{x}_0) \neq 0$.*

Proof. Using the chain rule, we differentiate both sides of $F(x, y, z) = k$ with respect to x_j, treating $x_j = x_j$ as a function of x_j, and x_i as a function of x_j and other variables as constants, we get $\dfrac{\partial F}{\partial x_j}\dfrac{\partial x_j}{\partial x_j} + \dfrac{\partial F}{\partial x_i}\dfrac{\partial x_i}{\partial x_j} = 0$ since all other $\dfrac{\partial x_w}{\partial x_j}$ terms are zero if x_w is constant. Also, $\dfrac{\partial x_j}{\partial x_j} = 1$. Solving for $\dfrac{\partial x_i}{\partial x_j}$ (at $\mathbf{x}_0$) gives us that $\dfrac{\partial x_i}{\partial x_j}(\mathbf{x}_0) = -\dfrac{F_{x_j}(\mathbf{x}_0)}{F_{x_i}(\mathbf{x}_0)}$. □

The conditions for which the independent variable is a function of the other variables locally is discussed in the Implicit Function Theorem. First, we should establish a few more preliminary differentiation laws, however.

Theorem 11.13. *Let $f, g : U \to \mathbb{R}^m$, where U is open in $\mathbb{R}^n$, f and g are differentiable at $\boldsymbol{p} \in U$, and $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_m(\boldsymbol{x}))$, and $g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), ..., g_m(\boldsymbol{x}))$. Then:*
(a) Sum rule: $D(\alpha f + \beta g)(\boldsymbol{p}) = \alpha Df(\boldsymbol{p}) + \beta Dg(\boldsymbol{p})$.
(b) Dot product rule: $D(f \cdot g)(\boldsymbol{p}) = g(\boldsymbol{p})Df(\boldsymbol{p}) + f(\boldsymbol{p})Dg(\boldsymbol{p})$.

Proof. (a) We note that $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{(\alpha f + \beta g)(\mathbf{p}+\mathbf{h}) - (\alpha f + \beta g)(\mathbf{p}) - \alpha Df(\mathbf{p})\mathbf{h} - \beta Dg(\mathbf{p})\mathbf{h}}{|\mathbf{h}|} = \lim\limits_{\mathbf{h}\to\mathbf{0}} \alpha \dfrac{f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - Df(\mathbf{p})\mathbf{h}}{|\mathbf{h}|} + \lim\limits_{\mathbf{h}\to\mathbf{0}} \beta \dfrac{g(\mathbf{p}+\mathbf{h}) - g(\mathbf{p}) - Dg(\mathbf{p})\mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$, which implies the desired result.

(b) We take $\lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{f(\mathbf{p}+\mathbf{h}) \cdot g(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) \cdot g(\mathbf{p}) - g(\mathbf{p})Df(\mathbf{p})\mathbf{h} - f(\mathbf{p})Dg(\mathbf{p})\mathbf{h}}{|\mathbf{h}|}$
$= \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{1}{|\mathbf{h}|} g(\mathbf{p}+\mathbf{h}) \cdot (f(\mathbf{p}+\mathbf{h}) - f(\mathbf{p}) - Df(\mathbf{p})\mathbf{h}) + \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{1}{|\mathbf{h}|} f(\mathbf{p}) \cdot (g(\mathbf{p}+\mathbf{h}) - g(\mathbf{p}) - Dg(\mathbf{p})\mathbf{h})$
$+ \lim\limits_{\mathbf{h}\to\mathbf{0}} \dfrac{1}{|\mathbf{h}|} Df(\mathbf{p})\mathbf{h} \cdot (g(\mathbf{p}+\mathbf{h}) - g(\mathbf{p}))$, assuming that each of these limits exists. The first two

summands are limits equal to zero since g, f are continuous and differentiable at $\mathbf{p}$. The third is also zero since $|\frac{1}{|\mathbf{h}|}Df(\mathbf{p})\mathbf{h}\cdot(g(\mathbf{p}+\mathbf{h})-g(\mathbf{p}))| \leq \frac{|Df(\mathbf{p})||\mathbf{h}|}{|\mathbf{h}|}|g(\mathbf{p}+\mathbf{h})-g(\mathbf{p})|$ by the definition of operator norm and the Cauchy Schwarz Inequality, and since g is continuous at $\mathbf{p}$ it follows that $\lim_{\mathbf{h}\to\mathbf{0}}|g(\mathbf{p}+\mathbf{h})-g(\mathbf{p})| = 0$. Thus, $D(f\cdot g)(\mathbf{p}) = g(\mathbf{p})Df(\mathbf{p}) + f(\mathbf{p})Dg(\mathbf{p})$.

□

There are a variety of different restrictions we can place on curves that help us to refer to them. Some books do a good job of distinguishing between the parametrization for a curve and the curve (the image of that parametrization) itself. We are going to take the approach that in some contexts a curve could be referring to the parametrization function whose image is the curve, and in other contexts we might mean the image (or trace) of the curve (the curve itself and not the function generating the curve). This intentional ambiguity is because of notation commonly used for line integrals. We will use notations such as $\int_C \mathbf{F}\cdot\mathbf{dr}$ which we are not going to define yet. But the "C" part of that notation refers to a curve. We want to talk about points on the curve (which means points on its trace), but the integral is not determined by the curve itself. Its orientation is needed, which means we are really saying the integral is dependent on the parametrization for the curve, meaning that we are using the same symbol, C, to refer to both the parametrization and to its image. This is a nuisance, and is similar to the problem of vector functions sometimes having outputs that refer to points, and vectors sometimes referring to matrices which are row or column vectors.

Definition 81

Let $\mathbf{r} : D \to \mathbb{R}^n$ be a continuous function whose domain is an interval D. We refer to this function (or its image depending on context) as a *parametrized curve* or just a *curve* C. When we want to be clear that we are specifically referring to the image of a parametrized curve $\mathbf{r}$ and not the function $\mathbf{r}$ itself, we refer to $\mathbf{r}(D)$ as the *trace* of C. Likewise, when we wish to be clear that we are referring to the function $\mathbf{r}$ and not to its trace, we call $\mathbf{r}$ a *parametrization* for C. A parametrized curve whose domain is a closed interval is also called a *path*.

If the parametrized curve $\mathbf{r} : (a,b) \to \mathbb{R}^n$ has a trace which, if intersected with some $B_\epsilon(\mathbf{r}(t_0))$, is also the graph of a function $y = f_1(x)$ or $x = f_2(y)$ for differentiable functions f_1 or f_2 then the trace of the curve is *locally a differentiable function graph near* $\mathbf{r}(t_0)$, and a tangent line to the trace of $\mathbf{r}((a,b))$ exists. In the case where $x'(t_0) = 0$ the tangent line is vertical (but still exists even though $\frac{dy}{dx}$ does not).

There are multiple parametrizations for the same curve. For instance, the circle in example (a) can be traversed by $\mathbf{r}(t) =< a + R\cos(2t), b + R\sin(2t) >$, $0 \leq t \leq \pi$, which is a parametrization tracing out the same curve counterclockwise at twice the speed of the former parametrization. We would like to formalize notions of speed and acceleration of a particle moving along a parametrized path, motivating the following definitions.

Let $\mathbf{r}(t) =< x_1(t), x_2(t), ..., x_n(t) >$ be a parametrized curve defined on an interval I. We should observe that for each point $t_0 \in I$, $\lim\limits_{t\to t_0} \mathbf{r}(t) =< z_1, z_2, ..., z_n >$ if and only if $\lim\limits_{t\to t_0} x_i(t) = z_i$ for each $1 \le i \le n$ by Theorem 10.17. Similarly, we know that $\mathbf{r}'(t)$ exists if and only if $x_i'(t)$ exists for each i, in which case $\mathbf{r}(t) =< x_1'(t), x_2'(t), ..., x_n'(t) >$.

Example 11.1. *Find* $\lim\limits_{t\to 0} < \dfrac{\sin(3t)}{t}, \dfrac{e^t - 1 - t}{t}, \dfrac{\cos(t) - 1}{t^2} >$.

Solution. We can just take the limit of each coordinate and use L'Hospital's Rule. For the first coordinate we get $\lim\limits_{t\to 0} \dfrac{\sin(3t)}{t} = \lim\limits_{t\to 0} \dfrac{3\cos(3t)}{1} = 3$. In the second coordinate, $\lim\limits_{t\to 0} \dfrac{e^t - 1 - t}{t} = \lim\limits_{t\to 0} \dfrac{e^t - 1}{1} = 0$ and in the third coordinate $\lim\limits_{t\to 0} \dfrac{\cos(t) - 1}{t^2} = \lim\limits_{t\to 0} \dfrac{-\sin(t)}{2t} = \lim\limits_{t\to 0} \dfrac{-\cos(t)}{2} = -\dfrac{1}{2}$. Thus, the limit is $< 3, 0, -\dfrac{1}{2} >$.

□

Example 11.2. *Find the derivative of* $\boldsymbol{r}(t) =< t, e^t, t^3 + 1 >$ *at* $(0, 1, 1)$.

Solution. Setting the first coordinates equal to each other we see $t = 0$. $\mathbf{r}'(t) =< 1, e^t, 3t^2 >$. Setting $t = 0$ we see $\mathbf{r}'(0) =< 1, 1, 0 >$.

□

Theorem 11.14. *Let $F : U \to \mathbb{R}$ be a differentiable function, where U is an open subset of $\mathbb{R}^n$, and let $\boldsymbol{x}_0$ be a point on the graph S of $F(\boldsymbol{x}) = k$. Let $\boldsymbol{r}(t)$ be a differentiable parametrized curve whose trace is in S so that $\boldsymbol{r}(t_0) = \boldsymbol{x}_0$. Then:*

(a) $\boldsymbol{r}'(t_0)$ is orthogonal to $\nabla F(\boldsymbol{x}_0)$.

(b) If S is locally the graph of a differentiable function $x_n = f(x_1, x_2, ..., x_{n-1})$ then $\nabla F(\boldsymbol{x}_0)$ is normal to S at $\boldsymbol{x}_0$.

(c) If S is locally the graph of a differentiable function $x_n = f(x_1, x_2, ..., x_{n-1})$ then if for every differentiable parametrized curve $\boldsymbol{r}(t)$ whose trace is in S so that $\boldsymbol{r}(t_0) = \boldsymbol{x}_0$ it is true that $\boldsymbol{r}'(t_0)$ is orthogonal to a vector $\boldsymbol{v}$ then $\boldsymbol{v} = \lambda \nabla F(\boldsymbol{x}_0)$ for some number λ.

Proof. (a) Since $F(\mathbf{r}(t)) = k$, we can use the chain rule to differentiate both sides, leaving us with $\nabla F(\mathbf{r}(t_0)) \cdot \mathbf{r}'(t_0) = 0$. Thus, $\mathbf{r}'(t_0)$ is perpendicular to $\nabla F(\mathbf{x}_0)$.

(b) Next, assume S is locally the graph of a differentiable function $x_n = f(x_1, x_2, ..., x_{n-1})$. Let $\mathbf{x}_0 = (c_1, c_2, ..., c_n)$, and let $\mathbf{c} = (c_1, c_2, ..., c_{n-1})$, so $f(\mathbf{c}) = c_n$. Then choose $\epsilon > 0$ so that $B_\epsilon(\mathbf{x}_0) \cap S$ is the graph of f, a function which is differentiable on some domain $D = \{(x_1, x_2, ..., x_{n-1}) \in \mathbb{R}^{n-1} | (x_1, x_2, ..., x_n) \in B_\epsilon(\mathbf{x}_0) \cap S\}$. By Theorem 11.11, we know that $\mathbf{n} = (\nabla f(\mathbf{c}), -1)$ is normal to $B_\epsilon(\mathbf{x}_0) \cap S$ at $\mathbf{x}_0$. However, since there is no sequence of points in S which converges to $\mathbf{x}_0$ which is contained in the complement of $B_\epsilon(\mathbf{x}_0)$, this also means that $\mathbf{n}$ is normal to S at $\mathbf{x}_0$.

Since $F(\mathbf{x}) - k = f(x_1, x_2, ..., x_{n-1}) - x_n = 0$ on $B_\epsilon(\mathbf{x}_0) \cap S$ it follows that $\nabla F(\mathbf{x}_0) = (\nabla f(\mathbf{c}), -1) = \mathbf{n}$.

(c) Assume S is locally the graph of a differentiable function $x_n = f(x_1, x_2, ..., x_{n-1})$ on some rectangle $\prod\limits_{i=1}^{n-1} (x_i - \epsilon, x_i + \epsilon)$ for some $\epsilon > 0$. For each $1 \le i \le n - 1$ let

$\mathbf{r}_i(t) = (c_1, c_2, c_3, ..., c_{i-1}, c_i + t, c_{i+1}, ..., c_{n-1}, f((c_1, c_2, ..., c_{i-1}, c_i + t, c_{i+1}, ..., c_{n-1}))$ for $t \in (-\epsilon, \epsilon)$. Then $\mathbf{r}_i(t) \subset S$, $\mathbf{r}_i(0) = \mathbf{x}_0$, and $\mathbf{r}_i'(0) = (0, 0, ..., 0, 1, 0, 0, ..., 0, \nabla f(\mathbf{c}) \cdot \mathbf{e}_i) = (0, 0, ..., 0, 1, 0, ..., 0, f_{x_i}(\mathbf{c}))$. Since each $\mathbf{r}_i'(0)$ is perpendicular to $\mathbf{v} = (v_1, v_2, ..., v_n)$, it must follow that each $\mathbf{v} \cdot \mathbf{r}_i'(0) = 0$ and therefore $v_i = -v_n f_{x_i}(\mathbf{c})$. Hence,
$\mathbf{v} = (-v_n f_{x_1}(\mathbf{x}_0), -v_n f_{x_2}(\mathbf{x}_0), ..., -v_n f_{x_{n-1}}(\mathbf{x}_0), v_n) = -v_n(\nabla f(\mathbf{c}), -1)$, which means that $\mathbf{v} = -v_n \nabla F(\mathbf{x}_0)$.

□

Theorem 11.15. *The Mean Value Theorem for Real Valued Functions. Let $f : U \to \mathbb{R}$ be a differentiable function, where U is an open set in $\mathbb{R}^n$ which contains a line segment $L(\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{a} \neq \boldsymbol{b}$. Then there is a point $\boldsymbol{c} \in L(\boldsymbol{a}, \boldsymbol{b})$ which is not equal to $\boldsymbol{a}$ or $\boldsymbol{b}$ so that $\nabla f(\boldsymbol{c}) \cdot (\boldsymbol{b} - \boldsymbol{a}) = f(\boldsymbol{b}) - f(\boldsymbol{a})$.*

Proof. Define $g(t) = f(\mathbf{a} + t(\mathbf{b} - \mathbf{a}))$, where $t \in [0, 1]$. Then g is continuous on $[0, 1]$ and differentiable on $(0, 1)$ by the chain rule, so by the Mean Value Theorem (for one variable), there is a point $d \in (0, 1)$ so that $g'(d)(1 - 0) = g(1) - g(0) = f(\mathbf{b}) - f(\mathbf{a})$. By the chain rule, $g'(d) = \nabla f(\mathbf{a} + d(\mathbf{b} - \mathbf{a})) \cdot (\mathbf{b} - \mathbf{a}) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a})$, where $\mathbf{c} = \mathbf{a} + d(\mathbf{b} - \mathbf{a})$. □

Theorem 11.16. *Let U be a connected open subset of $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a differentiable function so that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ for each $\boldsymbol{x} \in U$. Then for some number k it is true that $f(\boldsymbol{x}) = k$ for all $\boldsymbol{x} \in U$.*

Proof. Let $\mathbf{a}, \mathbf{b} \in U$. Then by Theorem 10.47, there is a polygonal path $P(\mathbf{a}, \mathbf{b}) = L(\mathbf{a}, \mathbf{x}_1) \cup L(\mathbf{x}_1, \mathbf{x}_2) \cup ... \cup L(\mathbf{x}_{m-1}, \mathbf{b}) \subseteq U$. For any line segment $L(\mathbf{p}, \mathbf{q}) \subseteq U$ we can find some $c \in L(\mathbf{p}, \mathbf{q})$ so that $f(\mathbf{q}) - f(\mathbf{p}) = \nabla f(\mathbf{c}) \cdot (\mathbf{q} - \mathbf{p}) = 0$ by the Mean Value Theorem for Real Valued Functions, which means that $f(\mathbf{q}) = f(\mathbf{p})$. Thus, it follows that $f(\mathbf{a}) = f(\mathbf{x}_1) = ... = f(\mathbf{b})$. Hence, $f(\mathbf{x}) = f(\mathbf{a}) = k$ for all $\mathbf{x} \in U$.

□

Theorem 11.17. *The Mean Value Theorem for Vector Valued Functions. Let $f : U \to \mathbb{R}^m$ be a differentiable function, where U is an open set in $\mathbb{R}^n$ which contains a line segment $L(\boldsymbol{a}, \boldsymbol{b})$, where $\boldsymbol{a} \neq \boldsymbol{b}$, and let $\boldsymbol{v} \in \mathbb{R}^m$. Then there is a point $\boldsymbol{c} \in L(\boldsymbol{a}, \boldsymbol{b})$ which is not equal to $\boldsymbol{a}$ or $\boldsymbol{b}$ so that $\boldsymbol{v} \cdot Df(\boldsymbol{c})(\boldsymbol{b} - \boldsymbol{a}) = \boldsymbol{v} \cdot (f(\boldsymbol{b}) - f(\boldsymbol{a}))$.*

Proof. Define $g(t) = \mathbf{v} \cdot (f(\mathbf{a} + t(\mathbf{b} - \mathbf{a})))$, where $t \in [0, 1]$. Then g is continuous on $[0, 1]$ and differentiable on $(0, 1)$ by the chain rule, so by the Mean Value Theorem (for one variable), there is a point $d \in (0, 1)$ so that $g'(d)(1 - 0) = g(1) - g(0) = \mathbf{v} \cdot (f(\mathbf{b}) - f(\mathbf{a}))$. By the chain rule, $g'(d) = \mathbf{v} \cdot Df(\mathbf{c})(\mathbf{b} - \mathbf{a})$, where $\mathbf{c} = \mathbf{a} + d(\mathbf{b} - \mathbf{a})$, which completes the proof. □

Theorem 11.18. *Let $f : U \to \mathbb{R}^m$ be C^1, where U is an open set in $\mathbb{R}^n$, and let $K \subset U$ be compact. Let $L(\boldsymbol{x}, \boldsymbol{y}) \subset U$ and let $|Df(\boldsymbol{z})| \leq M$ for all $\boldsymbol{z} \in L(\boldsymbol{x}, \boldsymbol{y})$. Then $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq M|\boldsymbol{x} - \boldsymbol{y}|$.*

If K is a compact subset of U and $L(\boldsymbol{x}, \boldsymbol{y}) \subseteq K$, and $M = \max\limits_{\boldsymbol{x} \in K} \sqrt{m} \sum_{i=1}^{m} \sum_{j=1}^{n} |\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})|$, then $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq M|\boldsymbol{x} - \boldsymbol{y}|$.

Proof. First, let $L(\mathbf{x}, \mathbf{y}) \subset U$. If $f(\mathbf{x}) = f(\mathbf{y})$ then the result is immediate. Assume $f(\mathbf{x}) \neq f(\mathbf{y})$. By the Mean Value Theorem for Vector Valued Functions, using $\mathbf{v} = f(\mathbf{y}) - f(\mathbf{x})$ from the theorem statement, we can find a point $\mathbf{c} \in L(\mathbf{x}, \mathbf{y})$ so that $(f(\mathbf{y}) - f(\mathbf{x})) \cdot Df(\mathbf{c})(\mathbf{y} - \mathbf{x}) = |f(\mathbf{y}) - f(\mathbf{x})|^2$. Hence, $|f(\mathbf{y}) - f(\mathbf{x})| \leq \dfrac{|f(\mathbf{y}) - f(\mathbf{x})||Df(\mathbf{c})||\mathbf{y} - \mathbf{x}|}{|f(\mathbf{y}) - f(\mathbf{x})|} \leq M|\mathbf{x} - \mathbf{y}|$.

Next, assume $L(\mathbf{x}, \mathbf{y}) \subset K$, a compact subset of U. Since f is C^1 we know that $\sqrt{m} \sum_{i=1}^{m} \sum_{j=1}^{n} |\frac{\partial f_i}{\partial x_j}|$ is continuous and has a maximum value M on K by the Extreme Value Theorem. By Theorem 10.32, $|Df(\mathbf{c})| \leq \sqrt{m} \sum_{i=1}^{m} \sum_{j=1}^{n} |\frac{\partial f_i}{\partial x_j}(\mathbf{c})| \leq \max\limits_{\mathbf{x} \in K} \sqrt{m} \sum_{i=1}^{m} \sum_{j=1}^{n} |\frac{\partial f_i}{\partial x_j}(\mathbf{x})|$ on K, so $|f(\mathbf{x}) - f(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$. □

It is also helpful to have a higher variable form of Taylor's Theorem. Unfortunately, multivariable Taylor series are more cumbersome than single variable series, but they are still quite useful.

Definition 82

Let $f : U \to \mathbb{R}$ be a C^k differentiable function, where U is an open subset of $\mathbb{R}^n$ which contains the line segment $L(\mathbf{x}, \mathbf{x} + \mathbf{h})$. We define the *kth order differential* of f at $\mathbf{x}$ with displacement $\mathbf{h}$ to be

$$D^{(k)} f(\mathbf{x}, \mathbf{h}) = \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \dots \sum_{i_k=1}^{n} \frac{\partial^{(k)} f}{dx_{i_k} dx_{i_{k-1}} \dots dx_{i_1}}(\mathbf{x}) h_{i_k} h_{i_{k-1}} \dots h_{i_1}.$$

Notice that $D^{(1)} f(\mathbf{x}, \mathbf{h}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}) h_i = \nabla f(\mathbf{x}) \cdot \mathbf{h}$, which is the standard differential approximation *df* to the change in f. Also, notice that if f is C^k then $D^{(1)}(D^{(k-1)} f(\mathbf{x}, \mathbf{h}))(\mathbf{x}, \mathbf{h})$

$$= \sum_{i_k=1}^{n} \frac{\partial f}{\partial x_{i_k}} \Big(\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \dots \sum_{i_{k-1}=1}^{n} \frac{\partial^{(k-1)} f}{dx_{i_{k-1}} dx_{i_{k-2}} \dots dx_{i_1}}(\mathbf{x}) h_{i_{k-1}} h_{i_{k-2}} \dots h_{i_1} \Big) h_{i_k} = D^{(k)} f(\mathbf{x}, \mathbf{h}).$$

This allows us to generalize Taylor's Theorem to n variables as follows:

Theorem 11.19. *Taylor's Theorem for $\mathbb{R}^n$. Let $f : U \to \mathbb{R}$ be a C^{k+1} function, where U is open in $\mathbb{R}^n$ and $L(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h}) \subset U$. Then $f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \sum_{i=1}^{k} \frac{1}{i!} D^{(i)} f(\boldsymbol{x}, \boldsymbol{h}) + \frac{1}{(k+1)!} D^{(k+1)} f(\boldsymbol{c}, \boldsymbol{h})$ for some point $\boldsymbol{c} \in L(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h})$.*

Proof. Let $g(t) = f(\mathbf{x}+t\mathbf{h})$ on $t \in [0,1]$. Then $g'(t) = \triangledown f(\mathbf{x}+t\mathbf{h})\cdot\mathbf{h}$ and $g'(0) = D^{(1)}f(\mathbf{x},\mathbf{h})$. Likewise, $g''(0) = D^{(2)}f(\mathbf{x},\mathbf{h})$ and so on, and, more generally, $g^{(i)}(t) = D^{(i)}f(\mathbf{x}+t\mathbf{h},\mathbf{h})$. Since g has $k+1$st derivative continuous on $[0,1]$ we can use Taylor's Theorem in one variable to get $g(1) = g(0) + \sum_{i=1}^{k} \frac{g^{(i)}(0)}{i!}(1)^i + \frac{1}{(k+1)!}g^{(k+1)}(d)(1-0)^{k+1}$ for some point $d \in [0,1]$.

Thus, $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \sum_{i=1}^{k} \frac{1}{i!}D^{(i)}f(\mathbf{x},\mathbf{h}) + \frac{1}{(k+1)!}D^{(k+1)}f(\mathbf{c},\mathbf{h})$ for some point $\mathbf{c} = \mathbf{x}+d\mathbf{h} \in L(\mathbf{x},\mathbf{x}+\mathbf{h})$.

□

Definition 83

Let $f : U \to \mathbb{R}$ be a function, where U is an open subset of R^n and $\mathbf{p} \in U$. We say that $(\mathbf{p}, f(\mathbf{p}))$ is a *local maximum* for f if there is some $\epsilon > 0$ so that $f(\mathbf{x}) \leq f(\mathbf{p})$ for all $\mathbf{x} \in B_\epsilon(\mathbf{p})$. We say that $(\mathbf{p}, f(\mathbf{p}))$ is a *local minimum* for f if there is some $\epsilon > 0$ so that $f(\mathbf{x}) \geq f(\mathbf{p})$ for all $\mathbf{x} \in B_\epsilon(\mathbf{p})$. We say that $(\mathbf{p}, f(\mathbf{p}))$ is a *local extremum* for f if $(\mathbf{p}, f(\mathbf{p}))$ is a local maximum or a local minimum for f. We say $(\mathbf{p}, f(\mathbf{p}))$ is a *saddle point* for f if $\triangledown f(\mathbf{p}) = \mathbf{0}$, but $(\mathbf{p}, f(\mathbf{p}))$ is not a local extremum of f. We will refer to a point $\mathbf{p}$ where $\triangledown f(\mathbf{p}) = \mathbf{0}$ as a *critical point* for f.

Theorem 11.20. *Let $f : U \to \mathbb{R}$ be a function which is differentiable at $\boldsymbol{p} \in U$, where U is open in $\mathbb{R}^n$ and $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local extremum. Then $\triangledown f(\boldsymbol{p}) = \boldsymbol{0}$.*

Proof. Let i be an integer so that $1 \leq i \leq n$. Define $g(t) = f(\mathbf{p}+t\mathbf{e}_i)$. Then g takes on a local extremum at $t = 0$ since $g(0) = f(\mathbf{p})$. Since g is differentiable at zero (by the chain rule), it follows that $g'(0) = f_{x_i}(\mathbf{p}) = 0$. Since this is true for all $1 \leq i \leq n$ it follows that $\triangledown f(\mathbf{p}) = \mathbf{0}$. □

This is helpful in finding extrema but it is not sufficient by itself, because, just as with single variable functions, a multivariable function may have saddle points. For example, $f(x,y) = x^2 - y^2$ has both partial derivatives equal to zero at $(0,0)$ but $(0,0,0)$ is not a local extremum for this function.

We might say a point is a saddle point for f or of f or for the graph of f or of the graph of f, and all are acceptable, and it is quite normal to just refer to such a point as a saddle point if the function or surface it is a saddle point of is understood. We look to additional tests to identify whether a point is a local extremum once we have found all critical points of a function. We can generalize a form of the first derivative test and a form of the second derivative test to dimensions higher than one.

The following theorem generalizes the first derivative test.

Theorem 11.21. *Let $f : U \to \mathbb{R}$ be a differentiable function, where U is open in R^n and $\boldsymbol{p}$ is a critical point of f. Then the following are true:*

(a) The point $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local maximum for f if there is an $\epsilon > 0$ so that for all $0 < t < \epsilon$, for every unit vector $\boldsymbol{u} \in \mathbb{R}^n$, the directional derivative $D_{\boldsymbol{u}} f(\boldsymbol{p} + t\boldsymbol{u}) \leq 0$.

(b) The point $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local minimum for f if there is an $\epsilon > 0$ so that for all $0 < t < \epsilon$, for every unit vector $\boldsymbol{u} \in \mathbb{R}^n$, the directional derivative $D_{\boldsymbol{u}} f(\boldsymbol{p} + t\boldsymbol{u}) \geq 0$.

Proof. We prove (a) first. Let $\mathbf{x} \in B_\epsilon(\mathbf{p})$. Then we can choose a unit vector $\mathbf{u}$ and $t_0 \in (0, \epsilon)$ so that $\mathbf{p} + t_0\mathbf{u} = \mathbf{x}$ (specifically, $\mathbf{u} = \dfrac{\mathbf{x} - \mathbf{p}}{|\mathbf{x} - \mathbf{p}|}$, and $t_0 = |\mathbf{x} - \mathbf{p}|$). Define $g(t) = f(\mathbf{p} + t\mathbf{u})$. Then for all $0 < t < \epsilon$, we know that $g'(t) = \nabla f(\mathbf{p} + t\mathbf{u}) \cdot (\mathbf{u}) = D_{\mathbf{u}} f(\mathbf{p} + t\mathbf{u}) \leq 0$. Thus, g is non-increasing on $[0, \epsilon)$ which means that $g(t_0) \leq g(0)$. Therefore, $f(\mathbf{x}) \leq f(\mathbf{p})$, which means that $(\mathbf{p}, f(\mathbf{p}))$ is a local maximum for f.

The proof of (b) is similar, or just observe that by negating a function f satisfying the hypotheses of (b) we have a function $-f$ satisfying the hypotheses of (a), and if $(\mathbf{p}, f(\mathbf{p}))$ is a local maximum for $-f$ then $(\mathbf{p}, f(\mathbf{p}))$ is a local minimum for f. □

There is also a generalization of the second derivative test. A more convenient generalization of this test works in $\mathbb{R}^2$. Both are outlined in the theorem below.

Theorem 11.22. *Second Derivative Test for multivariable functions.*

(a) Let $f : U \to \mathbb{R}$, where f is C^2 on the open set U in $\mathbb{R}^n$. If $\nabla f(\boldsymbol{p}) = 0$ for some $\boldsymbol{p} \in U$ and $D_{\mathbf{u}} D_{\boldsymbol{u}} f(\boldsymbol{p}) > 0$ for all $\boldsymbol{u} \in \mathbb{R}^n$ so that $|\boldsymbol{u}| = 1$, then $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local minimum for f. If $D_{\mathbf{u}} D_{\boldsymbol{u}} f(\boldsymbol{p}) < 0$ for all $\boldsymbol{u} \in \mathbb{R}^n$ so that $|\boldsymbol{u}| = 1$, then $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local maximum for f. If there is some unit vector $\boldsymbol{u}$ so that $D_{\mathbf{u}} D_{\boldsymbol{u}} f(\boldsymbol{p}) < 0$ and another unit vector $\boldsymbol{v}$ so that $D_{\boldsymbol{v}} D_{\boldsymbol{v}} f(\boldsymbol{p}) > 0$ then $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a saddle point for f

(b) Let $f : U \to \mathbb{R}$, where f is C^2 on the open set U in $\mathbb{R}^2$. Let $\boldsymbol{p}$ be a critical point of f. Let $D = f_{xx} f_{yy} - (f_{xy})^2$. If $D(\boldsymbol{p}) > 0$ then $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local maximum for f if $f_{xx}(\boldsymbol{p}) < 0$ and $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local minimum for f if $f_{xx}(\boldsymbol{p}) > 0$. If $D(\boldsymbol{p}) < 0$ then $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a saddle point for f.

Proof. (a) Choose a $\delta > 0$ so that $\overline{B_\delta(\mathbf{p})} \subset U$. We define $g_{\mathbf{u}}(t) = f(\mathbf{p} + t\mathbf{u})$. Using the chain rule, have that $g''_{\mathbf{u}}(t) = D_{\mathbf{u}} D_{\mathbf{u}} f(\mathbf{p} + t\mathbf{u}) = D^{(2)} f(\mathbf{p} + t\mathbf{u}, \mathbf{u})$ (as discussed with the proof of Taylor's Theorem for $\mathbb{R}^n$), which is continuous since all second partial derivatives of f are continuous.

Setting $G(\mathbf{x}, \mathbf{u}) = D_{\mathbf{u}} D_{\mathbf{u}} f(\mathbf{x})$ (a function whose domain is in $\mathbb{R}^{2n}$ where $(\mathbf{x}, \mathbf{u})$ represents the vector whose first n coordinates are those of $\mathbf{x}$ and whose second n coordinates are those of $\mathbf{u}$), we note that this is a continuous function over all $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{x} \in \overline{B_\delta(\mathbf{p})}$. If we let $D = \overline{B_2(\mathbf{0})}$ in $\mathbb{R}^n$, the closed radius two ball about the origin, then $H = \{(\mathbf{x}, \mathbf{u}) | \mathbf{x} \in \overline{B_\delta(\mathbf{p})}$ and $\mathbf{u} \in D\}$ is a closed and bounded set in $\mathbb{R}^{2n}$. Thus, G is uniformly continuous on H by Theorem 10.35.

Since $C = \{(\mathbf{p}, \mathbf{u}) | |\mathbf{u}| = 1\}$ is closed and bounded, $|G(\mathbf{x}, \mathbf{u})|$ takes on a minimum value m on C by the Extreme Value Theorem. Assuming that $G(\mathbf{x}, \mathbf{u}) \neq 0$ on C, then $m > 0$. Hence, if G is positive on C then $G(\mathbf{p}, \mathbf{u}) \geq m$ on C, and if G is negative on all of C then $G(\mathbf{p}, \mathbf{u}) \leq -m$ on C.

Since G is uniformly continuous, we can find a $0 < \gamma < \delta$ so that if $|(\mathbf{x}, \mathbf{u}) - (\mathbf{y}, \mathbf{v})| < \gamma$ then $|G(\mathbf{x}, \mathbf{u}) - G(\mathbf{y}, \mathbf{v})| < \dfrac{m}{2}$. Thus, for all $\mathbf{x} \in B_\gamma(\mathbf{p})$, if $|\mathbf{u}| = 1$ then $|G(\mathbf{x}, \mathbf{u})| > \dfrac{m}{2}$.

By the second derivative test in one variable, $g_{\mathbf{u}}(t)$ takes on a strict local maximum at $t = 0$ if $g''_{\mathbf{u}}(0) = D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) < 0$ and a strict local minimum if $D_{\mathbf{u}}D_{\mathbf{u}}(\mathbf{p}) > 0$. Furthermore, if G is negative on C then for all $\mathbf{x} = \mathbf{p} + t\mathbf{u} \in B_{\gamma}(\mathbf{p})$ it follows that $g''_{\mathbf{u}}(t) = D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{x}) < -\dfrac{m}{2}$, which means that $g'_{\mathbf{u}}(t) < 0$ for each unit vector $\mathbf{u}$ and all $0 < t < \gamma$ (since the first derivative is decreasing on $(0, \gamma)$ and is zero at $t = 0$). Thus, by the Theorem 11.21, the point $(\mathbf{p}, f(\mathbf{p}))$ is a local maximum for f. Likewise, if G is positive on C then by a similar argument $(\mathbf{p}, f(\mathbf{p}))$ is a local minimum for f.

If there is some unit vector $\mathbf{u}$ so that $D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) < 0$ and another unit vector $\mathbf{v}$ so that $D_{\mathbf{v}}D_{\mathbf{v}}f(\mathbf{p}) > 0$ then this means that for sufficiently small values of $t > 0$ it is true that $g_{\mathbf{u}}(t) > g_{\mathbf{u}}(0)$ and $g_{\mathbf{v}}(t) < g_{\mathbf{u}}(0)$. Hence, there are points $\mathbf{p} + t\mathbf{u}$ and $\mathbf{p} + t\mathbf{v}$ which are arbitrarily close to $\mathbf{p}$ so that $f(\mathbf{p} + t\mathbf{u}) > f(\mathbf{p})$ and $f(\mathbf{p} + t\mathbf{v}) < f(\mathbf{p})$, so $(\mathbf{p}, f(\mathbf{p}))$ is a saddle point for f.

(b) Let $\mathbf{u} = (\Delta x, \Delta y) \in \mathbb{R}^2$ be a unit vector. We, again, define $g_{\mathbf{u}}(t) = f(\mathbf{p} + t\mathbf{u})$. Using the chain rule, have that $g'_{\mathbf{u}}(t) = \nabla f(\mathbf{p} + t\mathbf{u}) \cdot \mathbf{u} = f_x(\mathbf{p} + t\mathbf{u})\Delta x + f_y(\mathbf{p} + t\mathbf{u})\Delta y$. Thus, $g''_{\mathbf{u}}(t) = (f_{xx}(\mathbf{p} + t\mathbf{u})\Delta x + f_{yx}(\mathbf{p} + t\mathbf{u})\Delta y)\Delta x + (f_{xy}(\mathbf{p} + t\mathbf{u})\Delta x + f_{yy}(\mathbf{p} + t\mathbf{u})\Delta y)\Delta y$ $= f_{xx}(\mathbf{p} + t\mathbf{u})(\Delta x)^2 + 2f_{xy}(\mathbf{p} + t\mathbf{u})\Delta x\Delta y + f_{yy}(\mathbf{p} + t\mathbf{u})(\Delta y)^2$ by Clairaut's Theorem. Thus, $g''_{\mathbf{u}}(0) = D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) = f_{xx}(\mathbf{p})(\Delta x)^2 + 2f_{xy}(\mathbf{p})\Delta x\Delta y + f_{yy}(\mathbf{p})(\Delta y)^2$.

We recall that for a quadratic equation $at^2 + bt + c$, there are no real zeroes for this equation if $b^2 - 4ac < 0$, and there are two real zeroes for this equation if $b^2 - 4ac > 0$. This fact comes from the quadratic formula, which can be proven by completing the square (but we will assume this is known from an earlier algebra course). For now, we will assume that $\Delta y \neq 0$ (the proof is similar if $\Delta x \neq 0$).

For $t = \dfrac{\Delta x}{\Delta y}$ we can rewrite $g''_{\mathbf{u}}(0) = D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) = f_{xx}(\mathbf{p})(\Delta x)^2 + 2f_{xy}(\mathbf{p})\Delta x\Delta y + f_{yy}(\mathbf{p})(\Delta y)^2 = (\Delta y)^2[f_{xx}(\mathbf{p})t^2 + 2f_{xy}(\mathbf{p})t + f_{yy}(\mathbf{p})]$. Since $(\Delta y)^2 > 0$, the sign of $g''_{\mathbf{u}}(0)$ is the same as the sign of $f_{xx}(\mathbf{p})t^2 + 2f_{xy}(\mathbf{p})t + f_{yy}(\mathbf{p})$. However, this is a quadratic expression in t and it has two zeros if $4(f_{xy}(\mathbf{p}))^2 - 4f_{xx}(\mathbf{p})f_{yy}(\mathbf{p}) > 0$ and no zeroes if $4(f_{xy}(\mathbf{p}))^2 - 4f_{xx}(\mathbf{p})f_{yy}(\mathbf{p}) < 0$. Hence, if $D(\mathbf{p}) = f_{xx}(\mathbf{p})f_{yy}(\mathbf{p}) - (f_{xy}(\mathbf{p}))^2 > 0$ and $f_{xx}(\mathbf{p}) > 0$ then that means $f_{xx}(\mathbf{p})t^2 + 2f_{xy}(\mathbf{p})t + f_{yy}(\mathbf{p}) > 0$ for all t (since this equation has no zeroes and yields a positive value when $t = 0$ because $f_{xx}(\mathbf{p})$ and $f_{yy}(\mathbf{p})$ must have the same sign and be non-zero if $f_{xx}(\mathbf{p})f_{yy}(\mathbf{p}) - (f_{xy}(\mathbf{p}))^2 > 0$). Likewise, if $f_{xx}(\mathbf{p}) < 0$ then $f_{xx}(\mathbf{p})t^2 + 2f_{xy}(\mathbf{p})t + f_{yy}(\mathbf{p}) < 0$ for all t. Hence, if $D(\mathbf{p}) > 0$ and $f_{xx}(\mathbf{p}) > 0$ then $D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) > 0$ for all unit vectors $\mathbf{u}$, so by part (a) we know $(\mathbf{p}, f(\mathbf{p}))$ is a local minimum for f. Similarly, if $D(\mathbf{p}) > 0$ and $f_{xx}(\mathbf{p}) < 0$ then $D_{\mathbf{u}}D_{\mathbf{u}}f(\mathbf{p}) < 0$ for all unit vectors $\mathbf{u}$, so by part (a) we know $(\mathbf{p}, f(\mathbf{p}))$ is a local maximum for f.

In the case where $D(\mathbf{p}) < 0$, the quadratic equation given has two zeroes and is therefore negative for some value t_0 of t and positive for some value t_1 of t. If we wish to be more specific, by picking angles θ_0, θ_1 so that $\cot(\theta_0) = t_0$ and $\cot(\theta_1) = t_1$, we can set $\Delta x_0 = \cos(\theta_0)$, $\Delta y_0 = \sin(\theta_0)$, $\Delta x_1 = \cos(\theta_1)$, and $\Delta y_1 = \sin(\theta_1)$. Then $\mathbf{u}_0 = (\Delta x_0, \Delta y_0)$ and $\mathbf{u}_1 = (\Delta x_1, \Delta y_1)$ are unit vectors so that $D_{\mathbf{u}_0}D_{\mathbf{u}_0}f(\mathbf{p}) > 0$ and $D_{\mathbf{u}_1}D_{\mathbf{u}_1}f(\mathbf{p}) < 0$, so by part (a) again, we know that $(\mathbf{p}, f(\mathbf{p}))$ is a saddle point for f.

□

Example 11.3. *Find all local extrema and saddle points for* $f(x, y) = x^4 - 4xy + y^4 + 1$.

Solution. Setting partials to zero we have $f_x = 4x^3 - 4y = 0$ so $y = x^3$, and $f_y = -4x + 4y^3 = 0$ so $x = y^3$. Thus, $(0,0), (1,1)$, and $(-1,-1)$ are critical points. We then evaluate $f_{xx} = 12x^2$, $f_{xy} = -4$ and $f_{yy} = 12y^2$. Thus, if $D = f_{xx}f_{yy} - (f_{xy})^2$ then $D(0,0) = -16 < 0$ so there is a saddle point at $(0,0,1)$. Since $D(1,1) = 128 > 0$ and $f_{xx}(1,1) = 12 > 0$, f has a local minimum $(1,1,-1)$. Since $D(-1,-1) = 128 > 0$ and $f_{xx}(-1,-1) = 12 > 0$, f also has a local minimum $(-1,-1,-1)$. This function has no local maxima.

□

The Inverse Function Theorem and the Implicit Function Theorem are important for many theorems.

Theorem 11.23. *The Inverse Function Theorem. We use variable notation $x_1 = x, x_2 = y, x_3 = z$. Let $f = (f_1, f_2, ..., f_n) : U \to \mathbb{R}^n$ be a C^1 function so that* $\det Df(\boldsymbol{p}) \neq 0$*, where $\boldsymbol{p} \in U$, an open in $\mathbb{R}^n$. Then there is an $r > 0$ so that:*

(a) The function $\det[\frac{\partial f_i}{\partial x_j}(\boldsymbol{c}_i)]_{n\times n}$ *is a continuous function of the entries $\boldsymbol{c}_i$ for $1 \leq i \leq n$, and for some $m > 0$ it is true that* $|\det[\frac{\partial f_i}{\partial x_j}(\boldsymbol{c}_i)]_{n\times n}| > m$ *for all $\boldsymbol{c}_i \in \overline{B_r(\boldsymbol{p})}$.*

(b) f is one to one on $\overline{B_r(\boldsymbol{p})}$

(c) $f(B_r(\boldsymbol{p}))$ is open and f restricted to $B_r(\boldsymbol{p})$ has inverse f^{-1} which is continuous on $f(B_r(\boldsymbol{p}))$.

(d) If we restrict f to $B_r(\boldsymbol{p})$ then $f^{-1} = (f_1^{-1}, f_2^{-1}, ..., f_n^{-1}) : f(B_r(\boldsymbol{p})) \to \mathbb{R}^n$ is C^1

(e) If we restrict f to $B_r(\boldsymbol{p})$ then f^{-1} is C^1 on $f(B_r(\boldsymbol{p}))$, and $Df^{-1}(f(\boldsymbol{x})) = (Df(\boldsymbol{x}))^{-1}$ for all $\boldsymbol{x} \in B_r(\boldsymbol{p})$

Proof. (a) The determinant is a sum of constants times products of continuous functions of the entries over the n-fold Cartesian product of U in $\mathbb{R}^{n^2}$, and is thus continuous.

Since $\det[\frac{\partial f_i}{\partial x_j}(\mathbf{p})]_{n\times n} = \det Df(\mathbf{p}) \neq 0$, and $\det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}$ is continuous at $(\mathbf{p}, \mathbf{p}, ..., \mathbf{p})$ (the vector with entries in $\mathbf{p}$ listed n times), we can find a $\delta > 0$ so that $\overline{B_\delta(\mathbf{p}, \mathbf{p}, ..., \mathbf{p})} \subseteq U \times U \times U ... \times U$ and $|\det Df(\mathbf{p}) - \det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}| < \frac{|\det Df(\mathbf{p})|}{2}$ for all $(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n) \in B_\delta(\mathbf{p}, \mathbf{p}, ..., \mathbf{p})$. Thus, $|\det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}| > \frac{|\det Df(\mathbf{p})|}{2}$ for all $(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n) \in B_\delta(\mathbf{p}, \mathbf{p}, ..., \mathbf{p})$ in $\mathbb{R}^{n^2}$. Choosing $r \in (0, \frac{\delta}{\sqrt{n}})$ we notice that if $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n \in \overline{B_r(\mathbf{p})}$ in $\mathbb{R}^n$ then $|(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n) - (\mathbf{p}, \mathbf{p}, ..., \mathbf{p})| = \sqrt{\sum_{i=1}^{n} |\mathbf{c}_i - \mathbf{p}|^2} < \sqrt{n}\frac{\delta}{\sqrt{n}} = \delta$, so $|\det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}| > \frac{|\det Df(\mathbf{p})|}{2} = m$.

(b) Let $\mathbf{y}, \mathbf{z} \in \overline{B_r(\mathbf{p})}$, where $\mathbf{y} = (y_1, y_2, .., y_n)$ and $\mathbf{z} = (z_1, z_2, .., z_n)$. Since each f_i is continuous, by the Mean Value Theorem for Real Valued Functions, for each $1 \leq i \leq n$ we can find $\mathbf{c}_i \in L(\mathbf{y}, \mathbf{z}) \subset \overline{B_r(\mathbf{p})}$ so that $\nabla f_i(\mathbf{c}_i) \cdot (\mathbf{z} - \mathbf{y}) = f_i(\mathbf{z}) - f_i(\mathbf{y})$. Suppose $f_i(\mathbf{z}) - f_i(\mathbf{y}) = 0$ for each $1 \leq i \leq n$. Choosing a $\mathbf{c}_i$ satisfying $\nabla f_i(\mathbf{c}_i) \cdot (\mathbf{z} - \mathbf{y}) = 0$ for each $1 \leq i \leq n$, gives a system of n equations in n variables $(z_i - y_i)$ whose coefficient matrix has a non-zero determinant $\det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}$ and therefore the system has a unique solution by Cramer's Rule.

By part (a) the determinant of the coefficient matrix is non-zero, so the the only solution is $z_i = y_i$ for each i, meaning that if $f(\mathbf{z}) = f(\mathbf{y})$ then $\mathbf{z} = \mathbf{y}$, so f is one to one on $B_r(\mathbf{p})$.

(c) Let $\mathbf{w} \in B_r(\mathbf{p})$. Since f is one to one on $\overline{B_r(\mathbf{p})}$, the function $g : \overline{B_r(\mathbf{p})} \to \mathbb{R}$ defined by $g(\mathbf{x}) = |f(\mathbf{w}) - f(\mathbf{x})|$ is positive on $\partial(B_r(\mathbf{p}))$. Since $\partial(B_r(\mathbf{p}))$ is closed and bounded, there is a least value l of g on $\partial(B_r(\mathbf{p}))$. In other words, for all $\mathbf{x} \in \partial(B_r(\mathbf{p}))$, the distance between $f(\mathbf{x})$ and $f(\mathbf{w})$ is at least l.

Let $\mathbf{q} \in B_{\frac{l}{2}}(f(\mathbf{w}))$. We then define a function $h_\mathbf{q} : \overline{B_r(\mathbf{p})} \to \mathbb{R}$ by $h_\mathbf{q}(\mathbf{x}) = |\mathbf{q} - f(\mathbf{x})|$. Since $\overline{B_r(\mathbf{p})}$ is compact and $h_\mathbf{q}$ is continuous, is must follow from the Extreme Value Theorem that $h_\mathbf{q}$ takes on a minimum value at some point $\mathbf{z} \in \overline{B_r(\mathbf{p})}$. Since $h_\mathbf{q}(\mathbf{w}) = |\mathbf{q} - f(\mathbf{w})| < \dfrac{l}{2}$ and for every point $\mathbf{b}$ on the boundary of $\overline{B_r(\mathbf{p})}$ we know that $|f(\mathbf{b}) - f(\mathbf{w})| \geq l$, it is impossible for $\mathbf{z}$ to be a point on the boundary of $\overline{B_r(\mathbf{p})}$ since, by the Triangle Inequality, $h_\mathbf{q}(\mathbf{b}) = |f(\mathbf{b}) - \mathbf{q}| \geq |f(\mathbf{b}) - f(\mathbf{w})| - |f(\mathbf{w}) - \mathbf{q}| > \dfrac{l}{2}$. Thus, we know that $\mathbf{z} \in B_r(\mathbf{w})$. Since $h_\mathbf{q}$ has a minimum at $\mathbf{z}$, it also follows that $(h_\mathbf{q})^2$ has a minimum at $\mathbf{z}$. By Theorem 11.20, it must follow that $\nabla(h_\mathbf{q})^2(\mathbf{z}) = \mathbf{0}$ (since $\mathbf{z}$ is in the interior of the domain of $h^2_\mathbf{q}$).

Let $\mathbf{q} = (q_1, q_2, ..., q_n)$. Since $h^2_\mathbf{q}(\mathbf{x}) = \displaystyle\sum_{i=1}^{n}(f_i(\mathbf{x}) - q_i)^2$, by the chain rule we have $\nabla(h_\mathbf{q})^2(\mathbf{z}) = (\displaystyle\sum_{i=1}^{n} 2(f_i(\mathbf{z}) - q_i)f_{i_{x_1}}(\mathbf{z}), \sum_{i=1}^{n} 2(f_i(\mathbf{z}) - q_i)f_{i_{x_2}}(\mathbf{z}), ..., \sum_{i=1}^{n} 2(f_i(\mathbf{z}) - q_i)f_{i_{x_n}}(\mathbf{z})) = (0, 0, 0, ..., 0)$. This gives the system of equations $\displaystyle\sum_{i=1}^{n}(f_i(\mathbf{z}) - q_i)f_{i_{x_j}}(\mathbf{z}) = 0$ for $1 \leq j \leq n$. Hence, treating the $(f_i(\mathbf{z}) - q_i)$ terms as the variables, the coefficient matrix is $Df(\mathbf{z}) \neq 0$, so there is a unique solution by Cramer's rule, which must be $f_i(\mathbf{z}) - q_i = 0$ for all $1 \leq i \leq n$, which means that $f(\mathbf{z}) = \mathbf{q}$. Thus, $B_{\frac{l}{2}}(f(\mathbf{w})) \subseteq f(B_r(\mathbf{p}))$. Hence, every point of $f(B_r(\mathbf{p}))$ is contained in an open ball which is contained in $f(B_r(\mathbf{p}))$, which means that $f(B_r(\mathbf{p}))$ is open.

Next, we wish to show that f restricted to $B_r(\mathbf{p})$ has inverse f^{-1} which is continuous on $f(B_r(\mathbf{p}))$. Let U be an open set in $\mathbb{R}^n$. Then $(f^{-1})^{-1}(U) = f(U) = f(U \cap B_r(\mathbf{p}))$ since f is one to one. For each point $\mathbf{x} \in U \cap B_r(\mathbf{p})$ we can find an $r_\mathbf{x} > 0$ by the argument above so that $B_{r_\mathbf{x}}(\mathbf{x}) \subset (U \cap B_r(\mathbf{p}))$ and $f(B_{r_\mathbf{x}}(\mathbf{x}))$ is open. Thus, $f(U) = \displaystyle\bigcup_{\mathbf{x} \in U \cap B_r(\mathbf{p})} f(B_{r_\mathbf{x}}(\mathbf{x}))$, which is open. Thus, f^{-1} is continuous.

(d) Let $f(\mathbf{x}_0) = \mathbf{y}_0 \in f(B_r(\mathbf{p}))$ and choose $R > 0$ so that $B_R(\mathbf{y}_0) \subset f(B_r(\mathbf{p}))$. Let $0 < |t| < R$. For each $1 \leq k \leq n$ there is a unique $\mathbf{x}_{(k,t)} \in B_r(\mathbf{p})$ so that $f(\mathbf{x}_{(k,t)}) = \mathbf{y}_0 + t\mathbf{e}_k$. By the Mean Value Theorem for Real Valued Functions we can find, for each $1 \leq i \leq n$ some $\mathbf{c}_i \in L(\mathbf{x}_0, \mathbf{x}_{(k,t)})$ so that $\dfrac{\nabla f_i(\mathbf{c}_i) \cdot (\mathbf{x}_{(k,t)} - \mathbf{x}_0)}{t} = \dfrac{f_i(\mathbf{x}_{(k,t)}) - f_i(\mathbf{x}_0)}{t}$, which is equal to zero if $i \neq k$ and equal to one if $i = k$. This creates a system of equations with coefficient matrix determinant $\det[\dfrac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n \times n}$ which is non-zero. By Cramer's Rule, for each $1 \leq m \leq n$, we can then solve for mth component of $\dfrac{\mathbf{x}_{(k,t)} - \mathbf{x}_0}{t}$, which is $\dfrac{f_m^{-1}(\mathbf{y}_0 + t\mathbf{e}_k) - f_m^{-1}(\mathbf{y}_0)}{t}$, which is equal to a ratio of determinants whose denominators are non-zero and whose matrix entries vary continuously with t. More specifically, $\dfrac{f_m^{-1}(\mathbf{y}_0 + t\mathbf{e}_k) - f_m^{-1}(\mathbf{y}_0)}{t} = \dfrac{\det A_m}{\det[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n \times n}}$, where A_m is the matrix obtained by replacing the mth column of $[\frac{\partial f_i}{\partial x_j}(\mathbf{c}_i)]_{n\times n}$, by $\mathbf{e}_m$.

As $t \to 0$, the points $\mathbf{c}_i \to \mathbf{x}_0$ by part (c) since f^{-1} is continuous. Thus, if we take the limit $\lim\limits_{t\to 0} \dfrac{f_m^{-1}(\mathbf{y}_0 + t\mathbf{e}_k) - f_m^{-1}(\mathbf{y}_0)}{t} = \dfrac{\partial f_m^{-1}(\mathbf{y}_0)}{\partial y_k} = \dfrac{\det B_m(\mathbf{x}_0)}{\det Df(\mathbf{x}_0)}$, where $B_m(\mathbf{x}_0)$ is obtained by replacing the mth column of $Df(\mathbf{x}_0)$ by $\mathbf{e}_m$.

Let $B_m(\mathbf{x})$ be the matrix obtained by replacing the mth column of $Df(\mathbf{x})$ by $\mathbf{e}_m$. The ratio $\dfrac{\det B_m(\mathbf{x})}{\det Df(\mathbf{x})} = \dfrac{\partial f_m^{-1}(\mathbf{y})}{\partial y_k}$, where $\mathbf{y} = f(\mathbf{x})$ varies continuously with $\mathbf{y}$ since f^{-1} is continuous on $f(B_r(\mathbf{p}))$, and both $\det B_m(\mathbf{x})$ and $\det Df(\mathbf{x})$ are continuous functions of $\mathbf{x}$, where $\det Df(\mathbf{x}) \neq 0$ on $f(B_r(\mathbf{p}))$. Thus, each of the partial derivatives of f^{-1} is continuous.

(e) The derivative of the identity function $g(\mathbf{x}) = \mathbf{x}$ on $\mathbb{R}^n$ is the identity matrix I_n consisting of one entries on the diagonal and zero entries everywhere else. This can be seen by simply taking the partial derivatives directly. Since we know f^{-1} is C^1 on $f(B_r(\mathbf{p}))$, for each $\mathbf{x} \in f(B_r(\mathbf{p}))$, by the chain rule we know that $D(f \circ f^{-1})(f(\mathbf{x})) = I_n = Df(f^{-1}(f(\mathbf{x}_0))Df^{-1}(f(\mathbf{x}_0))$. Hence, it follows that $Df^{-1}(f(\mathbf{x}_0))$ and $Df(\mathbf{x}_0)$ are inverses of each other (by Theorem 14.12).

□

The following is a minor corollary to the Inverse Function Theorem that can be helpful for deciding when topological properties are preserved. This is a sufficiently direct application of the Inverse Function Theorem that when we quote this result we may just say "by The Inverse Function Theorem."

Theorem 11.24. *Let $\phi : U \to \mathbb{R}^n$ be C^1 with $\Delta_\phi \neq 0$ on U, an open set in $\mathbb{R}^n$. Let V be an open subset of U. Then $\phi(V)$ is open. Furthermore, if ϕ is one to one then ϕ is a homeomorphism from U to $\phi(U)$.*

Proof. Let V be an open subset of U. Let $\phi(\mathbf{p}) \in \phi(V)$ for some $\mathbf{p} \in V$. By the Inverse Function Theorem there is some $\epsilon > 0$ so that $B_\epsilon(\mathbf{p}) \subseteq V$ and $\phi(B_\epsilon)$ is an open set containing $\phi(\mathbf{x})$ which is contained in $\phi(U)$. Hence, $\phi(V)$ is open.

If ϕ is one to one, since we know ϕ is C^1 and therefore continuous, and $(\phi^{-1})^{-1}(V) = \phi(V)$ is open for every open $V \subseteq U$, we know that ϕ^{-1} is continuous. Every function is onto its own image, so the function $\phi : U \to \phi(U)$ is a homeomorphism. □

Presenting the Implicit Function Theorem's most general form may not immediately make sense to some readers, so we plan to proceed along two approaches. First, we we present a proof that parallels methods of Courant's *Introduction to Calculus and Analysis (volume 2)*, which proves the theorem for two variables, in which context it is easy to discuss. Then we will present generalizations based on the argument included in Wade's *Introduction to Analysis* to higher dimensions. The hope is that by the time we prove the most general version the previous discussion will help us to explain what it means.

Theorem 11.25. *The Implicit Function Theorem for two variable real valued functions. Let $F : U \to \mathbb{R}$ be a C^1 function, where U is open in $\mathbb{R}^2$ and $(x_0, y_0) = \boldsymbol{p} \in U$ so that $F(x_0, y_0) = 0$ and $F_y(x_0, y_0) \neq 0$. Then there is a rectangle $[x_0-a, x_0+a]\times[y_0-b, y_0+b] \subset U$*

(where $a, b > 0$*) so that for each* $x \in [x_0 - a, x_0 + a]$ *there is exactly one* $y = f(x) \in [y_0 - b, y_0 + b]$ *so that* $F(x, f(x)) = 0$*. Furthermore,* $f(x_0) = y_0$ *and the function* f *is* C^1*, and* $f'(x) = -\dfrac{F_x}{F_y}(x, f(x))$*.*

Proof. We will assume $F_y(x_0, y_0) = 2m > 0$ (the argument if $m < 0$ is similar). Since F_y is continuous we can find $a_1, b > 0$ which are small enough so that $F_y > m$ on $[x_0 - a_1, x_0 + a_1] \times [y_0 - b, y_0 + b] \subset U$. Since $F_y > m$ it follows that $F(x, y)$ is an increasing function in the variable y for each $x \in [x_0 - a_1, x_0 + a_1]$, so there is at most one y so that $F(x, y) = 0$ for any given $x \in [x_0 - a_1, x_0 + a_1]$. By the Extreme Value Theorem, since F_x is continuous on a number $[x_0 - a_1, x_0 + a_1] \times [y_0 - b, y_0 + b]$, we can find a number M which exceeds $|F_x|$ on $[x_0 - a_1, x_0 + a_1] \times [y_0 - b, y_0 + b]$.

For each $x \in [x_0 - a_1, x_0 + a_1]$, by the Mean Value Theorem we can find c so that $F_x(c, y_0)(c - y_0) = F(x, y_0) - F(x_0, y_0) = F(x, y_0)$. Thus, $|F(x, y_0)| < Ma_1$. Furthermore, also by the Mean Value Theorem, we can find c_x for each $x \in [x_0 - a_1, x_0 + a_1]$ so that $F(x, y_0 + b) = F(x, y_0 + b) - F(x, y_0) + F(x, y_0) = F_y(x, c_x)(b) + F(x, y_0) \geq mb - Ma_1$. Likewise, we can find $d_x \in (y + 0 - b, y_0)$ so that $F(x, y_0 - b) = F(x, y_0 - b) + F(x, y_0) - F(x, y_0) = -F_y(x, c_x)(b) - F(x, y_0) \geq -mb + Ma_1$. Replacing a_1 by a smaller positive number a so that $mb - M(a + \epsilon) > 0$ for some $0 < \epsilon < a$, we have that $F(x, y_0 + b) > 0$ and $F(x, y_0 - b) < 0$ for all $x \in [x_0 - a, x_0 + a]$. Hence, by the Intermediate Value Theorem, for each $x \in [x_0 - a - \epsilon, x_0 + a + \epsilon]$ there is $y = f(x) \in [y_0 - b, y_0 + b]$ so that $F(x, f(x)) = 0$. Since we know that $F(x_0, y_0) = 0$ it follows that $f(x_0) = y_0$.

Let $x \in [x_0 - a, x_0 + a]$. Let $|h| < \epsilon$ and set $k = f(x + h) - f(x)$. By the Mean Value Theorem for Real Valued Functions we can find a point $\mathbf{c} \in L((x, f(x), (x + h, f(x) + k))$ so that $\nabla F(\mathbf{c}) \cdot (h, k) = F(x, f(x)) - F(x + h, f(x + h)) = 0$. Hence, $F_x(\mathbf{c})h + F_y(\mathbf{c})k = 0$, so $\dfrac{k}{h} = \dfrac{f(x + h) - f(x)}{h} = -\dfrac{F_x}{F_y}(\mathbf{c})$. Hence, $\displaystyle\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = -\frac{F_x}{F_y}(x, f(x))$. Since this limit exists, f is differentiable and therefore continuous. Since f is continuous, $-\dfrac{F_x}{F_y}(x, f(x))$ is continuous, and therefore $y = f(x)$ is C^1 on $[x_0 - a, x_0 + a] \times [y_0 - b, y_0 + b]$.

□

To help see what this theorem means, we look at a circle $x^2 + y^2 = 1$. The graph of this relation is not a function. We can see that y is not a function of x)nor is x a function of y since the vertical line test is failed (there is more than one y value for each x value apart from those at the left and right ends of the circle). We can write this as $F(x, y) = x^2 + y^2 - 1 = 0$. Then $F_y = 2y \neq 0$ unless $y = 0$. That means that if $y \neq 0$ then we can find a rectangle containing (x, y) on which, locally (ignoring the graph of F outside the small rectangle), $y = f(x)$ is a function of x. At $y = 0$ this is not true, and on the graph of this circle we can see that no matter how small a rectangle we take about $(1, 0)$ or $(-1, 0)$ the resulting piece of the graph of the circle would still fail the vertical line test (so y would not be a function of x). Furthermore, since F has a slope that varies continuously with x and y, the derivative of $f'(x)$ would be continuous over this small rectangle.

Note that this theorem does not tell us what the function f is, only that it exists, which is frequently important for us to know even when we cannot determine f.

Definition 84

Let $F : \mathbb{R}^m \to \mathbb{R}^n$ where $m = n + k$, where $k \in \mathbb{N}$. If $x_{s_1}, ..., x_{s_r}$ are variables of F and $r \leq n$ then we denote $\dfrac{\partial(f_{s_1}, f_{s_2}, ..., f_{s_r})}{\partial(x_{s_1}, ..., x_{s_r})} = \det[\dfrac{\partial f_{s_i}}{\partial x_{s_j}}]_{r\times r}$. We use the notation $(\mathbf{x}, \mathbf{t})$ to denote the vector in $\mathbb{R}^{n+k}$ whose first n coordinates are the entries of vector $\mathbf{x} \in \mathbb{R}^n$ and whose last k entries are those of vector $\mathbf{t} \in R^k$. It is understood that if we write the dimension of a Euclidean space as a sum of positive integers in this way and an element of the space as a pair of vectors then the first vector has a number of coordinates equal to the first integer listed in the sum and the second has a number of coordinates equal to the second integer in the sum.

Theorem 11.26. *The Implicit Function Theorem. Let $F = (F_1, F_2, ..., F_n) : U \to \mathbb{R}^n$ be C^1, where U is an open subset of R^{n+k} which contains the point $(\boldsymbol{x}_0, \boldsymbol{t}_0)$, and let $F(\boldsymbol{x}_0, \boldsymbol{t}_0) = \boldsymbol{0}$, where $\dfrac{\partial(F_1, F_2, ..., F_n)}{\partial(x_1, x_2..., x_n)}(\boldsymbol{x}_0, \boldsymbol{t}_0) \neq 0$. Then there is an open set $B_r(\boldsymbol{t}_0)$ in $\mathbb{R}^k$ so that for each $\boldsymbol{t} \in B_r(\boldsymbol{t}_0)$ there is a unique $\boldsymbol{x} = g(\boldsymbol{t}) \in \mathbb{R}^n$ so that $F(g(\boldsymbol{t}), \boldsymbol{t}) = \boldsymbol{0}$. Furthermore this function g is C^1 on $B_r(\boldsymbol{t}_0)$ and $g(\boldsymbol{t}_0) = \boldsymbol{x}_0$.*

Proof. We begin by defining $G(\mathbf{x}, \mathbf{t}) = (F(\mathbf{x}, \mathbf{t}), \mathbf{t})$ on U. Then the domain and range of G are in $\mathbb{R}^{n+k}$ and $\det DG(\mathbf{x}_0, \mathbf{t}_0) =$

$$\begin{vmatrix}
\frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & ... & \frac{\partial F_1}{\partial x_n} & \frac{\partial F_1}{\partial t_1} & \frac{\partial F_1}{\partial t_2} & ... & \frac{\partial F_1}{\partial t_k} \\
\frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & ... & \frac{\partial F_2}{\partial x_n} & \frac{\partial F_2}{\partial t_1} & \frac{\partial F_2}{\partial t_2} & ... & \frac{\partial F_2}{\partial t_k} \\
... & & & & & & & \\
\frac{\partial F_n}{\partial x_1} & \frac{\partial F_n}{\partial x_2} & ... & \frac{\partial F_n}{\partial x_n} & \frac{\partial F_n}{\partial t_1} & \frac{\partial F_n}{\partial t_2} & ... & \frac{\partial F_n}{\partial t_k} \\
0 & 0 & ... & 0 & 1 & 0 & ... & 0 \\
0 & 0 & ... & 0 & 0 & 1 & ... & 0 \\
... & & & & & & & \\
0 & 0 & ... & 0 & 0 & 0 & ... & 1
\end{vmatrix}(\mathbf{x}_0, \mathbf{t}_0) = \frac{\partial(F_1, F_2, ..., F_n)}{\partial(x_1, x_2..., x_n)}(\mathbf{x}_0, \mathbf{t}_0) \neq 0.$$

Thus, by the Inverse Function Theorem, there is some $R > 0$ so that (i) The function $\det[\dfrac{\partial G_i}{\partial x_j}(\mathbf{c}_i)]_{n+k\times n+k}$ is a continuous and non-zero function of the $\mathbf{c}_i$ entries for $\mathbf{c}_i \in B_R(\mathbf{x}_0, \mathbf{t}_0)$.

(ii) G is one to one on $B_R(\mathbf{x}_0, \mathbf{t}_0)$ and

(iii) $G(B_R(\mathbf{x}_0, \mathbf{t}_0))$ is open and $G^{-1}(\mathbf{y}, \mathbf{t}) = (G_1^{-1}(\mathbf{y}, \mathbf{t}), G_2^{-1}(\mathbf{y}, \mathbf{t}), ..., G_n^{-1}(\mathbf{y}, \mathbf{t}), \mathbf{t})$ is C^1 on $G(B_R(\mathbf{x}_0, \mathbf{t}_0))$. Note that the fact that the last k entries are those of $\mathbf{t}$ is not a consequence of the Inverse Function Theorem, but a consequence of the definition of G (the only point that can map to a point whose last k coordinates are those of $\mathbf{t}$ under G is a point whose last k coordinates are those of $\mathbf{t}$).

Since $G(B_R(\mathbf{x}_0, \mathbf{t}_0))$ is open, we can find some $r > 0$ so that $B_r(G(\mathbf{x}_0, \mathbf{t}_0)) \subset G(B_R(\mathbf{x}_0, \mathbf{t}_0))$, where $G(\mathbf{x}_0, \mathbf{t}_0) = (\mathbf{0}, \mathbf{t}_0)$ by construction, and of course $B_r(\mathbf{t}_0)$ is open in $\mathbb{R}^k$, and $\mathbf{t} \in B_r(\mathbf{t}_0)$ if and only if $(\mathbf{0}, \mathbf{t}) \in B_r(\mathbf{0}, \mathbf{t}_0)$.

We define $g(\mathbf{t}) = (G_1^{-1}(\mathbf{0}, \mathbf{t}), G_2^{-1}(\mathbf{0}, \mathbf{t}), ..., G_n^{-1}(\mathbf{0}, \mathbf{t}))$ for each $\mathbf{t} \in B_r(\mathbf{t}_0)$. Note that $g(\mathbf{t}_0) = \mathbf{x}_0$ since $G(\mathbf{x}_0, \mathbf{t}_0) = (\mathbf{0}, \mathbf{t}_0)$, so $g(\mathbf{t}_0)$ is the vector whose coordinates are the first n coordinates of the point mapping to $(\mathbf{0}, \mathbf{t}_0)$, which is $\mathbf{x}_0$.

Observe that $F(g(\mathbf{t}),\mathbf{t})$ is vector whose coordinates are the first n coordinates of the image of a point which is the inverse image of a point whose first n coordinates are zero, which means that $F(g(\mathbf{t}),\mathbf{t})=\mathbf{0}$ for all $\mathbf{t}\in B_r(\mathbf{t}_0)$.

Next, note that each partial $\dfrac{\partial g_i}{\partial t_j}(\mathbf{t})=\dfrac{\partial G_i^{-1}}{\partial t_j}(\mathbf{0},\mathbf{t})$ because $G_i^{-1}(\mathbf{0},\mathbf{t})=g_i(\mathbf{t})$ by definition of g, so $\lim\limits_{h\to 0}\dfrac{g_i(\mathbf{t}+h\mathbf{e}_{n+j})-g_i(\mathbf{t})}{h}=\lim\limits_{h\to 0}\dfrac{G_i^{-1}(\mathbf{0},\mathbf{t}+h\mathbf{e}_{n+j})-G_i^{-1}(\mathbf{0},\mathbf{t})}{h}$. We know $\dfrac{\partial G_i^{-1}}{\partial t_j}(\mathbf{0},\mathbf{t})$ varies continuously with $\mathbf{t}$ since G^{-1} is C^1, and therefore g is C^1 on $B_r(\mathbf{t}_0)$. Also, observe that $(g(\mathbf{t}),\mathbf{t})=G^{-1}(\mathbf{0},\mathbf{t})\in B_R(\mathbf{x}_0,\mathbf{t}_0)$ for all $\mathbf{t}\in B_r(\mathbf{t}_0)$.

Suppose $F(h(\mathbf{t}),\mathbf{t})=\mathbf{0}$ for some $\mathbf{t}\in B_r(\mathbf{t}_0)$. Then $G(h(\mathbf{t}),\mathbf{t})=(\mathbf{0},\mathbf{t})=G(g(\mathbf{t}),\mathbf{t})$. Since G is one to one on $B_R(\mathbf{x}_0,\mathbf{t}_0)$, it follows that $(h(\mathbf{t}),\mathbf{t})=(g(\mathbf{t}),\mathbf{t})$ which means that $h(\mathbf{t})=g(\mathbf{t})$.

Thus, we have established all points of the theorem and the proof is complete.

□

In some situations it seems to be less intuitive to have the g function have the first coordinates as outputs instead of the first coordinates. So, we also note that we could have done it the other way. Here is a minor restatement of the theorem.

Theorem 11.27. *Restatement of Implicit Function Theorem. Let $G=(G_1,G_2,...,G_k):U\to\mathbb{R}^k$ be C^1, where U is an open subset of R^{n+k} which contains the point $(\boldsymbol{x}_0,\boldsymbol{t}_0)$, and let $G(\boldsymbol{x}_0,\boldsymbol{t}_0)=\boldsymbol{0}$, where $\dfrac{\partial(G_1,G_2,...,G_k)}{\partial(t_1,t_2...,t_k)}(\boldsymbol{x}_0,\boldsymbol{t}_0)\neq 0$. Then there is an open set $B_r(\boldsymbol{x}_0)$ in $\mathbb{R}^n$ so that for each $\boldsymbol{x}\in B_r(\boldsymbol{x}_0)$ there is a unique $\boldsymbol{t}=g(\boldsymbol{x})\in\mathbb{R}^k$ so that $G(\boldsymbol{x},g(\boldsymbol{x}))=\boldsymbol{0}$. Furthermore this function g is C^1 on $B_r(\boldsymbol{x}_0)$ and $g(\boldsymbol{x}_0)=\boldsymbol{t}_0$.*

Proof. We apply the earlier statement of the Implicit Function Theorem, switching n and k and letting $G(\mathbf{x},\mathbf{t})=F(\mathbf{t},\mathbf{x})$ in the earlier statement.

□

An immediate consequence of the Implicit Function Theorem is the following:

Theorem 11.28. *Let $F:U\to\mathbb{R}$, where U is open in $\mathbb{R}^n$ containing $\boldsymbol{p}$, F is C^1 and S is the graph of $F(\boldsymbol{x})=k$. If $\nabla F(\boldsymbol{p})\neq 0$ on S then S is locally the graph of a C^1 function at $\boldsymbol{p}$. In particular, if $F_{x_i}(\boldsymbol{p})\neq 0$ for some i then S is locally the graph of a C^1 function $x_i=g(x_1,x_2,...,x_{i-1},x_{i+1},...,x_n)$ at $\boldsymbol{p}$.*

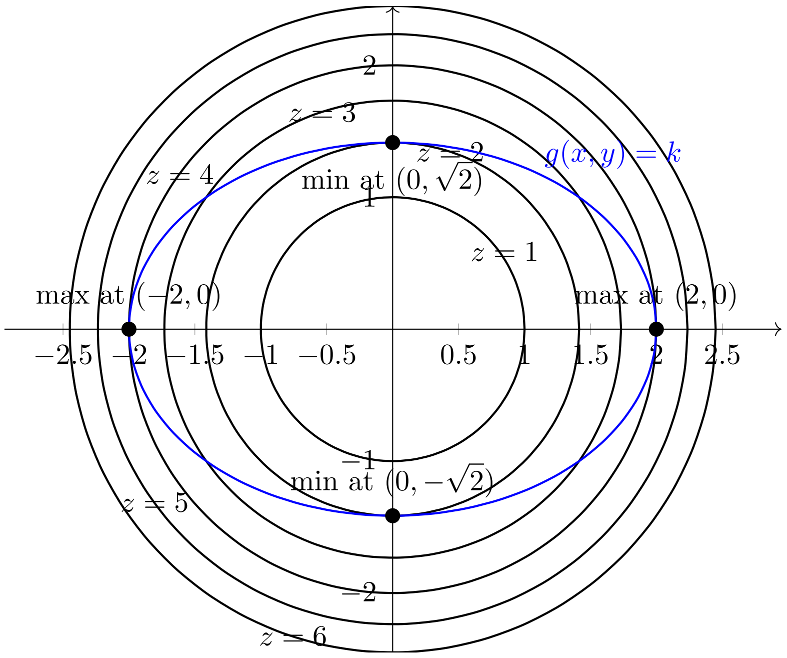
Proof. We will assume that $F_{x_n}(\mathbf{p})\neq 0$ (the other partials being non-zero leads to an argument that is similar, only changing variable names). Let $\mathbf{p}=(c_1,c_2,...,c_n)$ and let $\mathbf{c}=(c_1,c_2,...,c_{n-1})$. By the Implicit Function Theorem there is an open ball $B_r(\mathbf{c})$ so that for every $\mathbf{x}\in B_r(\mathbf{c})$ there is exactly one $x_n=g(\mathbf{x})$ so that $F(\mathbf{x},g(\mathbf{x}))=k$. The function g is C^1 on $B_r(\mathbf{c})$. Thus, $(B_r(\mathbf{c})\times\mathbb{R})\cap S$ is the graph of g on $B_r(\mathbf{c})$, and thus for any $0<\epsilon\leq r$ we know that $B_\epsilon(\mathbf{p})\cap S$ is the graph of g on the open set $D=\{(x_1,x_2,...,x_{n-1})\in\mathbb{R}^{n-1}|(x_1,x_2,...,x_{n-1},g(x_1,x_2,...,x_{n-1}))\in B_\epsilon(\mathbf{p})\cap S\}$.

□

Frequently, we wish to find extrema of a multivariable function $f : \mathbb{R}^n \to \mathbb{R}$ subject to a condition restricting the values in the domain to be considered, which may be designated as a constraint that $g(\mathbf{x}) = k$, for some $g : \mathbb{R}^n \to \mathbb{R}$. Assuming that f and g are both differentiable functions, it is helpful to use the method of Lagrange Multipliers to achieve this. It turns out that at any point $\mathbf{p}$ where there is a local extremum of f subject to constraint $g(\mathbf{x}) = k$, it will be true that $\nabla f(\mathbf{p}) = \lambda \nabla g(\mathbf{p})$ (for some number λ) assuming the constraint graph is differentiable, the reasons for which will be explained below. By solving this equation and the constraint equation as a system we are able to identify points at which an extremum could occur. In most problems in which Lagrange Multipliers are used we are looking for an absolute extremum rather than a local extremum (though this method does identify local extrema subject to the constraint as well). This usually requires an additional step of verifying that there actually is an absolute maximum and an absolute minimum at one of the relative extrema. If so, then by comparing the values of the function f at the points satisfying the aforementioned $\nabla f(\mathbf{p}) = \lambda \nabla g(\mathbf{p})$ system of equations, we can identify the largest of these function values as the absolute maximum value and the smallest as the absolute minimum value.

Before we prove the theorem about Lagrange Multipliers we mention that it is fairly easy to see the reason for the Lagrange Multiplier equation graphically for a two variable function. For a two variable function $z = f(x, y)$ if we use a contour map wherein we sketch the graphs of trace curves $f(x, y) = h$ (called contour lines) for many values of h, and on the same two dimensional graph we sketch the graph of $g(x, y) = k$, a constraint function, then if the graph of $g(x, y) = k$ passes through a contour line $f(x, y) = h_1$ then the graph $g(x, y) = k$ would also intersect nearby contour lines for heights h_0 smaller h_2 and larger than h_1 (on either side of the $f(x, y) = h_1$ contour line). Thus, the maximum of f subject to $g(x, y) = k$ could not occur at a point on the curve where the slope of $g(x, y) = k$ is different from the slope of $f(x, y) = h_1$ (since if the slopes are different the curves will cross through one another). However, the slopes of these curves are the same if and only if the gradients are parallel, which means that $\nabla f(\mathbf{p}) = \lambda \nabla g(\mathbf{p})$ at a point $\mathbf{p}$ where an extremum occurs.

Contour Map for $z = f(x, y)$ with Constraint Curve $g(x, y) = k$



Before we move further, we should be somewhat more precise about what we mean by a local maximum of a function subject to a constraint.

Definition 85

Let $f : U \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be differentiable functions, where U is open in $\mathbb{R}^n$. We say that a point $(\mathbf{p}, f(\mathbf{p}))$ (for some $\mathbf{p} \in U$) is a *local maximum for f subject to the constraint $g(\boldsymbol{x}) = k$* if $g(\mathbf{p}) = k$ and there is an $\epsilon > 0$ so that if $\mathbf{x} \in B_\epsilon(\mathbf{p})$ and $g(\mathbf{x}) = k$ then $f(\mathbf{x}) \leq f(\mathbf{p})$. Similarly, $(\mathbf{p}, f(\mathbf{p}))$ is a *local minimum for f subject to the constraint $g(\boldsymbol{x}) = k$* if $g(\mathbf{p}) = k$ and there is an $\epsilon > 0$ so that if $\mathbf{x} \in B_\epsilon(\mathbf{p})$ and $g(\mathbf{x}) = k$ then $f(\mathbf{x}) \geq f(\mathbf{p})$.

The following theorem outlines a proof for Lagrange's method of undetermined multipliers (the number λ being referred to as Lagrange's multipler) when only a single constraint it used.

Theorem 11.29. *Lagrange Multipliers with one constraint. Let $f : U \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be C^1 functions, where U is open in $\mathbb{R}^n$ and $\boldsymbol{p} \in U$ and $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local extremum for f with subject to the constraint $g(\boldsymbol{x}) = k$, where $\nabla g \neq \boldsymbol{0}$. Then for some number λ, it is true that $\nabla f(\boldsymbol{p}) = \lambda \nabla g(\boldsymbol{p})$.*

Proof. First, since $\nabla g \neq 0$, we know that the graph of $g(\mathbf{x}) = k$ is locally the graph of a differentiable function by Theorem 11.28. Thus, by Theorem 11.14 we know that $\nabla g(\mathbf{p})$ is normal to the graph of $g(\mathbf{x}) = k$ at $(\mathbf{p}, f(\mathbf{p}))$.

Let $\mathbf{r}(t)$ be a differentiable parametrized curve so that $g(\mathbf{r}(t)) = k$ and $\mathbf{r}(t_0) = \mathbf{p}$. Since $f(\mathbf{r}(t))$ is a one variable function that takes on a local maximum at $t = t_0$, it follows that $(f(\mathbf{r}(t))'(t_0) = 0$, so by the chain rule, $\triangledown f(\mathbf{r}(t_0)) \cdot \mathbf{r}'(t_0) = 0$. Thus, by Theorem 11.14, there is some λ so that $\triangledown f(\mathbf{p}) = \lambda \triangledown g(\mathbf{p})$.

□

Example 11.4. *Find the maximum and minimum values of $f(x, y, z) = xy + z^2$ subject to the constraint $x^2 + 2y^2 + 3z^2 = 12$.*

Solution. Both f and g are locally C^∞ functions at every point, so we don't need to worry about conditions for Lagrange multipliers not being satisfied anywhere. The ellipsoid constraint graph is closed and bounded so we know for certain that there will be a maximum and minimum because of the Extreme Value Theorem, and so we know that these extrema will occur where $\triangledown f = \lambda \triangle g$ and $g(x, y, z) = x^2 + 2y^2 + 3z^2$.

Thus, $f_x = \lambda g_x$, $f_y = \lambda g_x$ and $f_z = \lambda g_z$ and the constraint is satisfied where the extrema occur. This means the extrema must satisfy the following system of equations:

$y = \lambda(2x)$

$x = \lambda(4y)$

$2z = \lambda(6z)$

$x^2 + 2y^2 + 3z^2 = 12$

To solve this system it may be a good idea to try to isolate λ in the first three equations. If $x = 0$ then we cannot divide by $2x$ since that would be division by zero, so we consider multiple cases. If $x = 0$ then from the second equation we see that either $\lambda = 0$ or $y = 0$. It is not possible that $\lambda = 0$ because then $x = y = z = 0$ and that does not satisfy the constraint equation. Hence, if $x = 0$ then $y = 0$, so in the constraint we would have that $3z^2 = 12$, so $z = \pm 2$. Otherwise, $x \neq 0$ which means that $\lambda = \dfrac{y}{2x}$ and since $\lambda \neq 0$ we know $y \neq 0$ and thus $x \neq 0$ and $\lambda = \dfrac{x}{4y}$. From the third equation we know that either $z \neq 0$, in which case $\lambda = \dfrac{1}{3}$, or $z = 0$. Since $\lambda = \dfrac{y}{2x} = \dfrac{x}{4y}$ we have that $4y^2 = 2x^2$, so $x^2 = 2y^2$. If $z = 0$ then $4y^2 = 12$, so $y = \pm\sqrt{3}$ and $x = \pm\sqrt{6}$. If $z \neq 0$ then since $\lambda = \dfrac{1}{3}$ we know that $3y = 2x$ and $3x = 4y$, which is only possible if $x = y = 0$, which is a case we have already covered. Thus, the possible solutions to the system are $(0, 0, \pm 2)$, $(\pm\sqrt{6}, \pm\sqrt{3}, 0)$.

We test those solutions in the function f and compare the values to see which are largest and smallest to identify the extrema. This gives us:

$f(0, 0, \pm 2) = 4$

$f((\sqrt{6}, \sqrt{3}, 0) = \sqrt{18}$, meaning $(\sqrt{6}, \sqrt{3}, 0, \sqrt{18})$ is an absolute maximum for f subject to the constraint.

$f((-\sqrt{6}, -\sqrt{3}, 0) = \sqrt{18}$, meaning $(-\sqrt{6}, -\sqrt{3}, 0, \sqrt{18})$ is an absolute maximum for f subject to the constraint.

$f((-\sqrt{6}, \sqrt{3}, 0) = -\sqrt{18}$, meaning $(-\sqrt{6}, \sqrt{3}, 0, -\sqrt{18})$ is an absolute minimum for f subject to the constraint.

$f((\sqrt{6}, -\sqrt{3}, 0) = -\sqrt{18}$, meaning $(\sqrt{6}, -\sqrt{3}, 0, -\sqrt{18})$ is an absolute minimum for f subject to the constraint.

The wording of the question did not ask for all the absolute extrema, but rather the maximum and minimum value. The maximum value of f subject to the constraint is $\sqrt{18}$ and the minimum value of f subject to the constraint is $-\sqrt{18}$.

□

In some cases it is not optimal to use Lagrange multipliers, and solving for one variable and plugging it into the equation for another is sufficient and possibly easier, but we sometimes make errors when we do so. For example, in the preceding example, we could have solved for $z^2 = \dfrac{12 - 2y^2 - x^2}{3}$, and plugged into f to get $f(x, y, z) = xy + \dfrac{12 - 2y^2 - x^2}{3} = h(x, y)$ and tried for find the extrema of the resulting two variable function. We could then have solved to find local extrema using the techniques of the preceding section. Solving for critical points we would have had:

$f_x = y - \frac{2}{3}x = 0.$

$f_y = x - \frac{4}{3}y = 0.$

This system's only solution is at $(0, 0)$ and we could have solved for $z = \pm 2$, but the corresponding points would not have been the absolute extrema. In fact, $h(x, y) = xy + \dfrac{12 - 2y^2 - x^2}{3}$ has no minimum value because it becomes arbitrarily negative along $y = 0$ as x becomes large, whereas x cannot become any larger than $\sqrt{12}$ along the original constraint. When substituting a solution of a constraint into an equation is it important to make sure that no information is lost. It is true that if $x^2 + 2y^2 + 3z^2 = 12$ then $f(x, y, z) = xy + \dfrac{12 - 2y^2 - x^2}{3}$ but the constraint $x^2 + 2y^2 + 3z^2 = 12$ tells us more than that. Simply determining where local extrema of $h(x, y) = xy + \dfrac{12 - 2y^2 - x^2}{3}$ might occur did not tell us where the local extrema of f subject to the constraint were because knowing that $f(x, y, z) = xy + \dfrac{12 - 2y^2 - x^2}{3}$ does not tell us that $f(x, y, z) = xy + z^2$, where $z^2 = \dfrac{12 - 2y^2 - x^2}{3}$. In other words, the substitution step was not reversible. We further observe that at $(\sqrt{6}, \sqrt{3}, 0)$ the variable z is not locally a function of x and y but rather y is locally a function of x and z and x is locally a function of y and z, so the methods of the preceding section are not applicable for determining a local extremum if z is used as the dependent variable. This is much like the case in single variable calculus where we have to be wary of the boundary points of an interval and test the end points separately. In this case, the ellipse on the boundary of the domain over which z is a differentiable function of x and y in the constraint surface is a place where the extrema might not show up at a critical point. Thus, we warn the reader that while it is often quicker to substitute information from a constraint into an equation that unless we are careful when we do so we may lose information that will cause us to fail to notice one of the absolute extrema. Be careful that the extrema are known to be points that would appear where z is a differentiable function of x and y in the constraint or that the points where this is not true are checked separately if you do such substitutions.

We sometimes want to add a second constraint for a three variable function and so on. We can add as many contraints as we wish if the number of constraints is smaller than the number of variables, and the pertinent partial Jacobian is non-zero where the extrema occur. For simplicity we will set g and h to zero instead of an arbitrary constant k (this does not reduce the generality of the theorem since by subtracting k from both sides of a

$g(x,y,z) = k$ constraint we get a $G(x,y,z) = g(x,y,z) - k = 0$ constraint), and we will just do the proof for the two constraints for a three variable function case. The argument can be extended to any number of constraints which is smaller than the number of variables, but the general argument looks a bit messy and it is easy to get lost in the subscripts, so we are not going to include it here.

Theorem 11.30. *Lagrange Multipliers with two constraints. Let $f : U \to \mathbb{R}$, $g : \mathbb{R}^3 \to \mathbb{R}$ and $h : \mathbb{R}^3 \to \mathbb{R}$ be C^1 functions, where U is open in $\mathbb{R}^3$ and $\boldsymbol{p} = (x_0, y_0, z_0) \in U$ and $(\boldsymbol{p}, f(\boldsymbol{p}))$ is a local extremum for f subject to the constraints $g(x,y,z) = 0$ and $h(x,y,z) = 0$, where $g_x(\boldsymbol{p})h_y(\boldsymbol{p}) - h_y(\boldsymbol{p})g_x(\boldsymbol{p}) \neq 0$.*

Then there are numbers λ and μ so that $\nabla f(\boldsymbol{p}) = \lambda \nabla g(\boldsymbol{p}) + \mu \nabla h(\boldsymbol{p})$.

Proof. By the Implicit Function Theorem, for some $\epsilon > 0$ it is the case that for all $x \in B_\epsilon(\mathbf{p})$ satisfying both constraints it is true that $x = x(z)$ and $y = y(z)$ for some C^1 functions $x(z), y(z)$.

Treating x and y as functions of z, we can consider $f(x(z), y(z), z)$ as a function of one variable and since this function takes on an extremum at $z = z_0$ we know $f(x(z), y(z), z)'(z_0) = 0$. Thus, we can use the chain rule to write (1) $f_z(\mathbf{p}) + f_x(\mathbf{p})\dfrac{\partial x}{\partial z}(z_0) + f_y(\mathbf{p})\dfrac{\partial y}{\partial z}(z_0) = 0$.

Differentiating the constraint functions with respect to z we obtain (2) $g_z + g_y\dfrac{\partial y}{\partial z} + g_x\dfrac{\partial x}{\partial z} = 0$ and (3) $h_z + h_y\dfrac{\partial y}{\partial z} + h_x\dfrac{\partial x}{\partial z} = 0$. This is true at every point, and in particular at $\mathbf{p}$.

By Cramer's Rule, since $g_x(\mathbf{p})h_y(\mathbf{p}) - h_y(\mathbf{p})g_x(\mathbf{p}) \neq 0$, we can find unique λ, μ so that the equations (4) $f_x(\mathbf{p}) + \lambda g_x(\mathbf{p}) + \mu h_x(\mathbf{p}) = 0$ and (5) $f_y(\mathbf{p}) + \lambda g_y(\mathbf{p}) + \mu h_y(\mathbf{p}) = 0$ are both true.

Adding λ times equation (2) plus μ times equation (3) to equation (1) at point $\mathbf{p}$ yields $(f_z(\mathbf{p}) + \lambda g_z(\mathbf{p}) + \mu h_z(\mathbf{p})) + (f_x(\mathbf{p}) + \lambda g_x(\mathbf{p}) + \mu h_x(\mathbf{p}))\dfrac{\partial x}{\partial z}(z_0) + (f_y(\mathbf{p}) + \lambda g_y(\mathbf{p}) + \mu h_y(\mathbf{p}))\dfrac{\partial y}{\partial z}(z_0) = 0$. Thus, by (4) and (5) it must follow that this simplifies to (6) $f_z(\mathbf{p}) + \lambda g_z(\mathbf{p}) + \mu h_z(\mathbf{p}) = 0$.

Having established (4), (5) and (6), the result follows.

□

Example 11.5. *Find the absolute extrema of $f(x,y,z) = xyz$ subject to the constraints $g(x,y,z) = x^2 + y^2 = 8$ and $h(x,y,z) = x + y + z = 0$.*

Solution. The intersection of the two constraint surfaces is a regular simple closed curve which is closed and bounded, so we know that there will be a maximum and a minimum at a point where $\nabla f = \lambda \nabla g + \mu \nabla h$ and the constraints are satisfied. This gives us the following equations:

$yz = \lambda(2x) + \mu(1)$
$xz = \lambda(2y) + \mu(1)$
$xy = \lambda(0) + \mu(1)$
$x^2 + y^2 = 8$

$x+y+z=0$

If x, y or z are zero then f is zero, which cannot be the minimum or maximum of f on the curve of intersection of the two constraints since we can see that we can get positive and negative values for f on the constraint curve simply by looking at the graph and seeing that there are points where two coordinates are positive and one is negative or two are negative and one is positive. This means that we can divide by any variable without dividing by zero.

We note that $\mu = xy$ by the third equation. Plugging this into the first two equations and solving for λ gives us that $\lambda = \frac{yz-xy}{2x} = \frac{xz-xy}{2y}$. Multiplying by $2xy$ gives $y^2z - xy^2 = x^2z - x^2y$. Using the last equation we have $z = -x-y$, so $y^2(-x-y) - xy^2 = x^2(-x-y) - x^2y$, so $y^2(2x+y) = x^2(2y+x)$ and $(x^2+y^2)(2y+x-2x-y) = 0$. Since $x^2+y^2 = 8$ from the fourth equation we have that $8(y-x) = 0$ which tells us that $x = y$. From the fourth equation this gives us that either $x = 2 = y$ so $z = -4$, or $x = -2 = y$ and $z = 4$. Testing these values we get $f(2,2,-4) = -16$, so $(2,2,-4,-16)$ is the absolute minimum and $f(-2,-2,4) = 16$, so $(-2,-2,4,16)$ is the absolute maximum of f subject to the two constraints. □

There are other methods for finding an extremum subject to a smooth constraint curve, such as parametrizing the curve. In the next example, we contrast both methods for a particular function.

Example 11.6. *Find the absolute extrema of $f(x,y) = 2x+3y$ subject to the constraint $x^2+y^2 = 1$.*

Solution. Here the constraint curve is just the unit circle. Proceeding with Lagrange multipliers in the usual way we have the system:

$2 = \lambda(2x)$

$3 = \lambda(2y)$

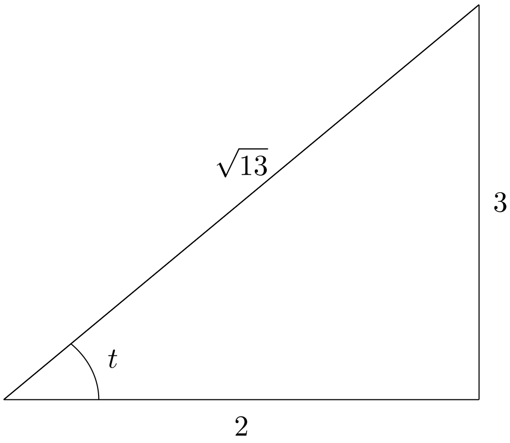
$x^2+y^2 = 1$

Hence, we know that $x, y \neq 0$ and solving for λ gives us $\lambda = \frac{1}{x} = \frac{3}{2y}$, so $2y = 3x$ and $y = \frac{3}{2}x$. Substituting this into the third equation gives us $x^2 + \frac{9}{4}x^2 = 1$, so $x^2 = \frac{4}{13}$ and $x = \pm\frac{2}{\sqrt{13}}$. If $x = \frac{2}{\sqrt{13}}$ then $y = \frac{3}{\sqrt{13}}$, and if $x = -\frac{2}{\sqrt{13}}$ then $y = -\frac{3}{\sqrt{13}}$. Since rationalized denominators are traditionally considered more simplified, we would say that the points at which extrema may occur are $(\frac{2\sqrt{13}}{13}, \frac{3\sqrt{13}}{13})$ and $(\frac{-2\sqrt{13}}{13}, \frac{-3\sqrt{13}}{13})$. Plugging into f we get that $f(\frac{2\sqrt{13}}{13}, \frac{3\sqrt{13}}{13}) = \frac{4\sqrt{13}}{13} + \frac{9\sqrt{13}}{13} = 13\frac{\sqrt{13}}{13} = \sqrt{13}$ and $f(\frac{-2\sqrt{13}}{13}, \frac{-3\sqrt{13}}{13}) = -\sqrt{13}$. Thus, f has an absolute maximum $(\frac{2\sqrt{13}}{13}, \frac{3\sqrt{13}}{13}, \sqrt{13})$ and an absolute minimum $(\frac{-2\sqrt{13}}{13}, \frac{-3\sqrt{13}}{13}, -\sqrt{13})$.

Next, we use the other method to find the extrema of the function subject to the constraint which we mentioned. We parametrize the unit circle as $\mathbf{r}(t) =< \cos(t), \sin(t) >$ over $[0, 2\pi]$ then we see that on the constraint curve we have $F(t) = f(\mathbf{r}(t)) = 2\cos(t) +$

$3\sin(t)$, which has derivative $F'(t) = 3\cos(t) - 2\sin(t)$. We then use the usual method for finding absolute extrema of a differentiable function of one variable on a closed interval. We check F at the end points of the parametrization (0 and 2π) and at the values of t where $F'(t) = 0$. We see $F(0) = F(2\pi) = 2$. Setting $3\cos(t) - 2\sin(t) = 0$ we have that $\tan(t) = \frac{3}{2}$, so $t = \tan^{-1}(\frac{3}{2})$ or $t = \pi + \tan^{-1}(\frac{3}{2})$. Drawing a triangle we see that if $t = \tan^{-1}(\frac{3}{2})$ then we have the following triangle sides from the Pythagorean Theorem.

□



From this we see that $\sin(t) = \frac{3}{\sqrt{13}}$ and $\cos(t) = \frac{2}{\sqrt{13}}$. In the case where $t = \pi + \tan^{-1}(\frac{3}{2})$, the sine and cosine are negated, so plugging into F we get the same values as before for the absolute extrema.

In this case (and many others) the method of Lagrange multipliers is probably a little easier, but there are instances where thinking of the right parametrization is probable better than the usual Lagrange multiplier process. Usually, if in doubt, it is best to assume Lagrange multipliers will probably be the easier process and look for a parmatrization only after finding that the generated equation system does not seem tractable.

When we are trying to find absolute extrema over a piecewise smooth curve, we look at each piece separately, and the hypotheses for Lagrange multipliers usually only work at points other than the end points of the piece decomposition, so the ends are tested separately. If parametrizations are used, we likewise parametrize each piece and use the ends of the piece decomposition intervals as end points to test as well.

Here is an example with a square constraint. Each side of the square is part of the graph of a differentiable function, but the corners are not points where the curve is differentiable.

Example 11.7. *Find the absolute extrema of $f(x, y) = 2x + 3y$ subject to the constraint curve consisting of the square whose sides are contained in $x = \pm 1$ and $y = \pm 1$.*

Solution. Here the constraint curve is a square. We have not satisfied the criteria for the constraint curve at the ends of the sides of the square, but if an extremum occurs inside the sides of the square other than at the corners, it should show up with Lagrange multipliers. This means we will have to check the corners of the square. At other points, we have either

$x = \pm 1$ or $y = \pm 1$. Thus, depending on which side of the square we are in, we could have a system looking like one of these:

$2 = \lambda(1)$ or $2 = \lambda(0)$
$3 = \lambda(0)$ or $3 = \lambda(1)$
$x = \pm 1$ or $y = \pm 1$

This system has no solutions so there are no extrema on the constraint except at the corners (where the Lagrange multipliers theorem is not applicable). Since the Extreme Value Theorem guarantees that f does have absolute extrema over the closed and bounded square we just test the corners to identify the absolute extrema. At $(1, 1)$ we get $f(1, 1) = 5$ is the absolute maximum value, and at $(-1, -1)$ we get $f(-1, -1) = -5$ is the absolute minimum value.

□

Now, in this example the location of the absolute extrema was fairly clear from the outset. There are quite a few problems like that, but in many cases (probably most) it is hard to simply see the value that will lead to an extremum.

Not all constrained domains over which we may wish to find an extremum are graphs of a level curve or surface of regular curve. For example, we might want to find the extrema of a function over the closed set bounded by an ellipse in the plane (or an ellipsoid in three dimensional Euclidean space). In such cases we find critical points in the interior of the set in question since an absolute extremum on the interior of a domain must also be a local extremum, and then use Lagrange multipliers (or parametrizations) to find the absolute extrema on the constraint. We compare the values of the function at the critical points in the interior and the absolute extrema on the boundary to find the absolute extrema over the set.

Example 11.8. *Find the absolute extrema of $f(x, y) = x^2 + y^2 + xy^2$ over $D = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 \leq 1\}$.*

Solution. We need both local extrema and extrema on the boundary curve. So, we take

$f_x = 2x + y^2 = 0$
$f_y = 2y + 2xy = 0$

Thus $2x = -y^2$, so $2y - y^3 = 0$ and so $y(2 - y^2) = 0$ which means that $y = 0$ or $y = \pm\sqrt{2}$. If $y = 0$ then $x = 0$ and if $y = \pm\sqrt{2}$ then $x = -1$.

Only one of these critical points is inside the unit disk D, so we include $(0, 0)$ in our set of points to be tested.

We then use Lagrange multipliers to find the absolute extrema on the boundary of D, which is the constraint circle $x^2 + y^2 = 1$. We could parametrize the circle, but most likely the Lagrange multiplier process will be more efficient. This gives us:

$2x + y^2 = \lambda(2x)$
$2y + 2xy = \lambda(2y)$
$x^2 + y^2 = 1$

If $x = 0$ then $y = \pm 1$. If $y = 0$ then $x = \pm 1$. Otherwise we can divide by $2x$ and $2y$ to give $\lambda = 1 + \frac{y^2}{2x} = 1 + x$, so $2x^2 = y^2$ and therefore $3x^2 = 1$ from the third equation, so $x = \pm\frac{1}{\sqrt{3}}$. Thus, $y = \pm\sqrt{\frac{2}{3}}$. Checking each of these possibilities we see that $f(0, 0) = 0$, $f(0, \pm 1) = 1$, $f(\pm 1, 0) = 1$, $f(\frac{\sqrt{3}}{3}, \pm\frac{\sqrt{6}}{3}) = 1 + \frac{1}{\sqrt{3}}$ and $f(-\frac{\sqrt{3}}{3}, \pm\frac{\sqrt{6}}{3}) = 1 - \frac{1}{\sqrt{3}}$. Hence, the absolute minimum is $(0, 0, 0)$ and the absolute maxima are $(\frac{\sqrt{3}}{3}, \pm\frac{\sqrt{6}}{3}, 1 + \frac{\sqrt{3}}{3})$.

□

It is natural to ask "should we test points that are local extrema of the function to be optimized which occur on the boundary constraint rather than on the interior of the domain over which we are looking for extrema?" The answer is that it does no harm to do so but it is unnecessary since those points would show up as potential extrema subject to the boundary constraint if they were points at which an extremum could occur.

Exercises:

Exercise 11.1. *Let V be a convex open subset of $\mathbb{R}^n$ and let $f : V \to \mathbb{R}^m$ be C^1. Prove that if $|Df(\boldsymbol{x})|$ is bounded then f is uniformly continuous.*

Exercise 11.2. *Let I be a non-empty open interval and let $f : I \to R$ be differentiable on I. If $f(I)$ is contained in the boundary of some open ball $B_r(\boldsymbol{0})$ about the origin, then prove that $f(t)$ and $f'(t)$ are orthogonal for all $t \in I$.*

Exercise 11.3. *Let $B = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 < 1\}$ and let $f : B \to \mathbb{R}^2$ be a C^1 one to one function so that $\triangle_f \neq 0$ on B. Let M be a connected subset of $f(B)$. Then $f^{-1}(M)$ is connected.*

Exercise 11.4. *Prove that there are functions $u(x, y), v(x, y), w(x, y)$, and an $r > 0$ such that u, v, w are continuously differentiable and satisfy the equations*

$$\begin{aligned}
&u^5 + xv^2 - y + w = 3\\
&v^5 + yu^2 - x + w = 3\\
&w^2 + y^5 - x^4 = 4
\end{aligned}$$

on $B_r(1, 1)$, so that $u(1, 1) = 1$, $v(1, 1) = 1$, and $w(1, 1) = 2$.

Exercise 11.5. *Give the second order Taylor polynomial based at the origin in the direction $< h, k >$, and write the formula for the remainder for $f(x, y) = e^{xy}$.*

Exercise 11.6. *Let $f(u, v) = (uv, u^2 + v^2)$ be defined on the portion of the first quadrant of the uv-plane with $v > u$. Find the derivative of f^{-1} at the point $(3, 10)$.*

Exercise 11.7. *Find the absolute extrema of $f(x, y, z) = xy$ over the compact solid bounded by the ellipsoid $x^2 + y^2 + z^2 = 18$.*

Exercise 11.8. *Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ and $g : \mathbb{R}^3 \to \mathbb{R}^4$ be defined by $f(x, y) = (x^2y, x+2y, 3y+z)$ and $g(u, v, w) = (v^2, 3u + v, 2w + 3v, u + w)$. Use the Chain Rule to find the derivative of $g \circ f$ at $(1, 2)$.*

Exercise 11.9. *Let $f(x,y,z,w):\mathbb{R}^4\to\mathbb{R}$ be C^2. Prove that $f_{xz}=f_{zx}$. More generally, if $g:\mathbb{R}^n\to\mathbb{R}$ and x_i,x_j are variables of the domain then $g_{x_ix_j}=g_{x_jx_i}$.*

Exercise 11.10. *Let $g:\mathbb{R}^n\to\mathbb{R}$ be a C^{n+2} function. Then for any finite sequence of integers $i_1,i_2,...,i_k\in\{1,2,...,n\}$, and any permutation (one to one correspondence) P of the order of these finitely many integers to give a re-ordering $P(i_i,P(i_2),...,P(i_k)$, it is always true that $g_{x_{i_1}x_{i_2}\cdots x_{i_k}}=g_{x_{P(i_1)}x_{P(i_2)}\cdots x_{P(i_k)}}$.*

Exercise 11.11. *Find the tangent plane to the surface $z^2=2z+2x+5xy+y^2$ at the point $(1,1,4)$.*

Exercise 11.12. *Taylor's series in $\mathbb{R}^n$.*

Let $f:U\to\mathbb{R}$ be a C^∞ function, where U is open in $\mathbb{R}^n$ and $L(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h})\subset U$. Then $f(\boldsymbol{x}+\boldsymbol{h})=f(\boldsymbol{x})+\sum_{i=1}^{\infty}\frac{1}{i!}D^{(i)}f(\boldsymbol{x},\boldsymbol{h})$.

Exercise 11.13. *True or false (assume sets are in $\mathbb{R}^n$, and give a brief justification for your answers):*

(a) The continuous image of a closed set is always closed.

(b) Every differentiable function is continuous.

(c) Every function whose partial derivatives exist at every point is continuous.

(d) Every function whose partial derivatives exist at every point is differentiable.

(e) Every function whose partial derivatives are continuous at every point of the domain of the function is differentiable.

(f) The continuous image of a closed bounded set is always closed.

(g) The boundary of a set is always closed.

(h) A function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is continuous if and only if each of its component functions is continuous.

(i) A function is differentiable if and only if each of its component functions is differentiable.

(j) Let $f:\mathbb{R}^2\to\mathbb{R}$. If the $\lim_{x\to0}f(ax,bx)=0$ for all $<a,b>\in\mathbb{R}^2$ then $\lim_{(x,y)\to(0,0)}f(x,y)=0$.

(k) Let $f:\mathbb{R}^2\to\mathbb{R}$. If the first partial derivatives of f exist and are continuous everywhere and the second partial derivatives exist at the point (a,b) then $f_{xy}(a,b)=f_{yx}(a,b)$.

(l) The continuous image of a connected set is always closed.

(m) The inverse image of a connected set under a continuous function is always connected.

(n) If the graph of a real valued function defined on a closed interval is closed and connected then it is also compact.

(o) The union of two compact sets is always compact.

(p) The intersection of two connected sets is always connected.

(q) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every open set $U \subset \mathbb{R}^n$ the set $f^{-1}(U)$ is open in E.

(r) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every closed set $A \subset \mathbb{R}^n$ the set $f^{-1}(A)$ is closed in E.

(s) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every compact set $K \subset \mathbb{R}^n$ the set $f^{-1}(K)$ is compact.

Solutions:

Solution to Exercise 11.1. *Let V be a convex open subset of $\mathbb{R}^n$ and let $f : V \to \mathbb{R}^m$ be C^1. Prove that if $|Df(\boldsymbol{x})|$ is bounded then f is Lipshcitz (and thus uniformly continuous) on V.*

Proof. Choose M so that $|Df(\mathbf{x})| \leq M$ on V. Let $\mathbf{x}, \mathbf{y} \in V$. Since V is convex, we know that $L(\mathbf{x}, \mathbf{y}) \subset V$. By the Mean Value Theorem for Vector Valued Functions we can find $\mathbf{c} \in L(\mathbf{x}, \mathbf{y})$ so that $(f(\mathbf{x}) - f(\mathbf{y})) \cdot Df(\mathbf{c})(\mathbf{x} - \mathbf{y}) = |f(\mathbf{x}) - f(\mathbf{y})|^2$. Thus, $|f(\mathbf{x}) - f(\mathbf{y})|^2 \leq |f(\mathbf{x}) - f(\mathbf{y})||Df(\mathbf{c})||\mathbf{x} - \mathbf{y}|$, so $|f(\mathbf{x}) - f(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$, so f is Lipschitz (and uniformly continuous) on V. □

Solution to Exercise 11.2. *Let I be a non-empty open interval and let $f : I \to \mathbb{R}^n$ be differentiable on I. If $f(I)$ is contained in the boundary of some open ball $B_r(\boldsymbol{0})$ about the origin, then prove that $f(t)$ and $f'(t)$ are orthogonal for all $t \in I$.*

Proof. Since $f(t) \cdot f(t) = r^2$ is constant, we can use the dot product rule to get that $2f'(t) \cdot f(t) = 0$, so $f'(t)$ is perpendicular to $f(t)$.

□

Solution to Exercise 11.3. *Let $B = \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 < 1\}$ and let $f : B \to \mathbb{R}^2$ be a C^1 one to one function so that $\triangle_f \neq 0$ on B. Let M be a connected subset of $f(B)$. Then $f^{-1}(M)$ is connected.*

Proof. By Theorem 11.24, f is a homeomorphism and f^{-1} is continuous. Since the continuous image of a connected set is connects, $f^{-1}(M)$ is connected. □

Solution to Exercise 11.4. *Prove that there are functions $u(x, y), v(x, y), w(x, y)$, and an $r > 0$ such that u, v, w are continuously differentiable and satisfy the equations*

$$\begin{aligned}
&u^5 + xv^2 - y + w = 3\\
&v^5 + yu^2 - x + w = 3\\
&w^2 + y^5 - x^4 = 4
\end{aligned}$$

on $B_r(1, 1)$, so that $u(1, 1) = 1$, $v(1, 1) = 1$, and $w(1, 1) = 2$.

Proof. Let $F(u, v, w, x, y) = (u^5 + xv^2 - y + w - 3, v^5 + yu^2 - x + w - 3, w^2 + y^5 - x^4 - 4)$. Note that $F(1, 1, 2, 1, 1) = (0, 0, 0)$ and that F is C^1 on all of $\mathbb{R}^5$. Also note that $\dfrac{\partial(F_1, F_2, F_3)}{\partial(u, v, w)} = \begin{vmatrix} 5u^4 & 2xv & 1 \\ 2uy & 5v^4 & 1 \\ 0 & 0 & 2w \end{vmatrix}$, so $\dfrac{\partial(F_1, F_2, F_3)}{\partial(u, v, w)}(1, 1, 2, 1, 1) = \begin{vmatrix} 5 & 2 & 1 \\ 2 & 5 & 1 \\ 0 & 0 & 4 \end{vmatrix} = 84 \neq 0.$

By The Implicit Function Theorem, there is an $r > 0$ and a unique C^1 function $g : B_r(0,0) \to \mathbb{R}^3$ so that $F(g(x,y),x,y) = (0,0,0)$ for all $(x,y) \in B_r(0,0)$, where $g(x,y) = (u(x,y), v(x,y), w(x,y))$ and thus u, v, w are C^1 on $B_r(0,0)$. Since $F(g(x,y),x,y) = (0,0,0)$, it follows that $u^5 + xv^2 - y + w = 3$, $v^5 + yu^2 - x + w = 3$, and $w^2 + y^5 - x^4 = 4$ for all $(x,y) \in B_r(0,0)$.

□

Solution to Exercise 11.5. *Give the second order Taylor polynomial based at the origin in the direction $< h,k >$, and write the formula for the remainder for $f(x,y) = e^{xy}$.*

Solution. The first and second order derivatives are $f_x = ye^{xy}$, $f_y = xe^{xy}$, $f_{xx} = y^2e^{xy}$, $f_{yy} = x^2e^{xy}$ and $f_{xy} = f_{yx} = e^{xy} + xye^{xy}$. The only first or second partial derivatives that are non-zero at $(0,0)$ are $f_xy(0,0) = f_{yx}(0,0) = 1$. Thus, the second order Taylor polynomial is $f(h,k) = 1 + 0(h) + 0(k) + 0(h^2) + 0(k^2) + 1(hk) + 1(kh) = 1 + \dfrac{2hk}{2!} = 1 + hk$. The third order derivatives are $f_{xxx} = y^3e^{xy}$, $f_{yyy} = x^3e^{xy}$, $f_{xxy} = 2ye^{xy} + xy^2e^{xy} = f_{xyx}$ and $f_{yyx} = 2xe^{xy} + x^2ye^{xy} = f_{yxy}$. Thus, the remainder is $\dfrac{1}{3!}e^{c_1c_2}(k^3c_1^3 + h^3c_2^3 + (4c_2 + 2c_1c_2^2)h^2k + (4c_1 + 2c_2c_1^2)hk^2)$ for some $c_1 \in (0,h), c_2 \in (0,k)$.

□

Solution to Exercise 11.6. *Let $f(u,v) = (uv, u^2 + v^2)$ be defined on the portion of the first quadrant of the uv-plane with $v > u$. Find the derivative of f^{-1} at the point $(3,10)$.*

Solution. We have $Df(u,v) = \begin{bmatrix} v & 2u \\ u & 2v \end{bmatrix}$. Since the determinant of the derivative matrix is non-zero and the function is one to one with a non-zero determinant, by the Inverse Function Theorem the function f^{-1} is C^1 and the derivative of f at $(3,10)$ is the inverse of Df at the inverse of $(3,10)$, which is $(1,3)$, where $Df(1,3) = \begin{bmatrix} 3 & 2 \\ 1 & 6 \end{bmatrix}$. The inverse of this matrix is $\dfrac{1}{16}\begin{bmatrix} 6 & -2 \\ -1 & 3 \end{bmatrix}$, which is the derivative of f^{-1} at $(3,10)$.

□

Solution to Exercise 11.7. *Find the absolute extrema of $f(x,y,z) = xyz$ over the compact solid E bounded by the ellipsoid $x^2 + y^2 + z^2 = 12$.*

Solution. Since E is compact, there is a maximum and a minimum value of f on E. The extrema could occur at the boundary or on the interior of a solid. To check the interior we set $f_x = yz = 0$ and $f_y = xz = 0$, $f_z = xy = 0$. Thus, the points where all partial derivatives are zero are those where two of the variables are zero. However, if any variable is zero then the function's value is zero, which is not the maximum or minimum value of the function on E.

Using Lagrange multipliers then, if an extremum occurs on the boundary, $f_x = yz = \lambda(2x)$, $f_y = xz = \lambda(2y)$, and $f_z = xy = \lambda(2z)$. Since the extrema do not occur when a variable is zero for this function, we can divide to get $\lambda = \frac{xz}{2y} = \frac{xy}{2z} = \frac{yz}{2x}$. Hence, $x^2 = y^2 = z^2 = 4$. Checking the points where this occurs, we have $f(2,2,2) = 8$, $f(2,-2,-2) = 8$, $f(-2,2,-2) = 8$, $f(-2,-2,2) = 8$, giving the absolute maxima, and $f(2,2,-2) = f(2,-2,2) = f(-2,2,2) = f(-2,-2,-2) = -8$, giving the absolute minima.

□

Solution to Exercise 11.8. *Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ and $g : \mathbb{R}^3 \to \mathbb{R}^4$ be defined by $f(x,y) = (x^2y, x+2y, 3y+x)$ and $g(u,v,w) = (v^2, 3u+v, 2w+3v, u+w)$. Use the Chain Rule to find the derivative of $g \circ f$ at $(1,2)$.*

Solution. By the Chain Rule, of $D(g \circ f)(1,2) = Dg(f(1,2))Df(1,2)$. We know $Dg = \begin{bmatrix} 0 & 2v & 0 \\ 3 & 1 & 0 \\ 0 & 3 & 2 \\ 1 & 0 & 1 \end{bmatrix}$ and $Df = \begin{bmatrix} 2xy & x^2 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$ and that $f(1,2) = (2,5,7)$. Hence, $D(g \circ f)(1,2) =$

$$Dg(f(1,2))Df(1,2) = \begin{bmatrix} 0 & 10 & 0 \\ 3 & 1 & 0 \\ 0 & 3 & 2 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 20 \\ 13 & 5 \\ 5 & 12 \\ 5 & 4 \end{bmatrix}.$$

□

Solution to Exercise 11.9. *Let $f(x,y,z,w) : \mathbb{R}^4 \to \mathbb{R}$ be C^2. Prove that $f_{xz} = f_{zx}$. More generally, if $g : \mathbb{R}^n \to \mathbb{R}$ is C^2 and x_i, x_j are variables of the domain then $g_{x_ix_j} = g_{x_jx_i}$.*

Proof. Applying Clairaut's Theorem, if y and w are held fixed then $f(x,y,z,w)$ is a function of two variables, namely x and z, so $f_{xz} = f_{zx}$ by Clairaut's Theorem.

In general, if $g : \mathbb{R}^n \to \mathbb{R}$ and x_i, x_j are variables of the domain then the partial derivatives with respect to these variables are found treating the other variables as fixed values. Thus means that g can be considered a two variable function for purposes of partial derivatives with respect to just the variables x_i, x_j, so $g_{x_ix_j} = g_{x_jx_i}$. □

Solution to Exercise 11.10. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a C^{n+2} function. Then for any finite sequence of integers $i_1, i_2, ..., i_k \in \{1,2,...,n\}$, and any permutation (one to one correspondence) P of the order of these finitely many integers to give a re-ordering $P(i_i, P(i_2), ..., P(i_k)$, it is always true that $g_{x_{i_1}x_{i_2}...x_{i_k}} = g_{x_{P(i_1)}x_{P(i_2)}...x_{P(i_k)}}$.*

Proof. Let $p_1 = P(i_1)$. By Exercise 11.9, we have $g_{x_{i_1}x_{i_2}..x_{p_1-1}x_{p_1}} = g_{x_{i_1}x_{i_2}..x_{p_1-2}x_{p_1}x_{p_1-1}}$, which means that $g_{x_{i_1}x_{i_2}..x_{p_1-1}x_{p_1}...x_{i_k}} = g_{x_{i_1}x_{i_2}..x_{p_1-2}x_{p_1}x_{p_1-1}...x_{i_k}}$. Hence, we can switch any adjacent two variables in order. By switching adjacent pairs until x_{p_1} is in the first position we see that $g_{x_{P(i_1)}x_{P(i_2)}...x_{P(i_k)}} = g_{i_1x_{P(i_2)}...x_{P(i_1)}...x_{P(i_k)}}$. Then by repeatedly switching adjacent pairs (we can move x_{i_2} into the second order derivative position and so on until we get that $g_{x_{i_1}x_{i_2}\dots x_{i_k}} = g_{x_{P(i_1)}x_{P(i_2)}\dots x_{P(i_k)}}$.

□

Solution to Exercise 11.11. *Find the tangent plane to the surface* $z^2 = 2z+2x+5xy+y^2$ *at the point* $(1,1,4)$.

Solution. Setting $F = 2z+2x+5xy+y^2-z^2$, we have $F_x = 2+5y$, $F_y = 5x+2y$, $F_z = 2-2z$, so $\nabla F(1,1,4) =< 7, 7, -6 >$. Tangent plane is $7(x-1)+7(y-1)-6(z-4) = 0$.

□

Solution to Exercise 11.12. *Taylor's series in* $\mathbb{R}^n$*. Let* $f : U \to \mathbb{R}$ *be a* C^∞ *function, where* U *is open in* $\mathbb{R}^n$ *and* $L(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) \subset U$*, and* $\lim\limits_{k\to\infty} \dfrac{1}{(k+1)!} \max\limits_{\boldsymbol{c}\in L(\boldsymbol{x},\boldsymbol{x_h})} D^{(k+1)}f(\boldsymbol{c},\boldsymbol{h}) = 0$. *Then* $f(\boldsymbol{x}+\boldsymbol{h}) = f(\boldsymbol{x}) + \sum\limits_{i=1}^{\infty} \dfrac{1}{i!} D^{(i)} f(\boldsymbol{x},\boldsymbol{h})$.

Proof. By Taylor's Theorem for multivariable functions, for every $k \in \mathbb{N}$ we know $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \sum\limits_{i=1}^{k} \dfrac{1}{i!} D^{(i)} f(\mathbf{x},\mathbf{h}) + \dfrac{1}{(k+1)!} D^{(k+1)} f(\mathbf{c},\mathbf{h})$ for some point $\mathbf{c} \in L(\mathbf{x},\mathbf{x}+\mathbf{h})$. Since $\lim\limits_{k\to\infty} \dfrac{1}{(k+1)!} \max\limits_{\mathbf{c}\in L(\mathbf{x},\mathbf{x_h})} D^{(k+1)} f(\mathbf{c},\mathbf{h}) = 0$, it follows that $\lim\limits_{k\to\infty} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \sum\limits_{i=1}^{k} \dfrac{1}{i!} D^{(i)} f(\mathbf{x},\mathbf{h}) = 0$, so $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \sum\limits_{i=1}^{\infty} \dfrac{1}{i!} D^{(i)} f(\mathbf{x},\mathbf{h})$. □

Solution to Exercise 11.13. *True or false (assume sets are in* $\mathbb{R}^n$*, and give a brief justification for your answers):*

(a) The continuous image of a closed set is always closed.

(b) Every differentiable function is continuous.

(c) Every function whose partial derivatives exist at every point is continuous.

(d) Every function whose partial derivatives exist at every point is differentiable.

(e) Every function whose partial derivatives are continuous at every point of the domain of the function is differentiable.

(f) The continuous image of a closed bounded set is always closed.

(g) The boundary of a set is always closed.

(h) A function from $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *is continuous if and only if each of its component functions is continuous.*

(i) A function is differentiable if and only if each of its component functions is differentiable.

(j) Let $f : \mathbb{R}^2 \to \mathbb{R}$*. If the* $\lim\limits_{x\to 0} f(ax, bx) = 0$ *for all* $< a, b > \in \mathbb{R}^2$ *then* $\lim\limits_{(x,y)\to(0,0)} f(x,y) = 0$.

(k) Let $f : \mathbb{R}^2 \to \mathbb{R}$*. If the first partial derivatives of* f *exist and are continuous everywhere and the second partial derivatives exist at the point* (a,b) *then* $f_{xy}(a,b) = f_{yx}(a,b)$.

(l) The continuous image of a connected set is always closed.

(m) The inverse image of a connected set under a continuous function is always connected.

(n) If the graph of a real valued function defined on a closed interval is closed and connected then it is also compact.

(o) The union of two compact sets is always compact.

(p) The intersection of two connected sets is always connected.

(q) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every open set $U \subset \mathbb{R}^n$ the set $f^{-1}(U)$ is open in E.

(r) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every closed set $A \subset \mathbb{R}^n$ the set $f^{-1}(A)$ is closed in E.

(s) A function $f : E \to \mathbb{R}^m$, where $E \subseteq \mathbb{R}^n$ is continuous if and only if for every compact set $K \subset \mathbb{R}^n$ the set $f^{-1}(K)$ is compact.

Solution. (a) False. The continuous image of a compact set is always compact and therefore closed but, we could, for instance take $f(x) = \dfrac{1}{x}$ on $(0, \infty)$, which takes $[1, \infty)$ (which is closed) to $(0, 1]$ (which is not).

(b) True by Theorem 11.3.

(c) False. A counterexample would be $f(x, y) = \dfrac{2x^2y}{x^4 + y^2}$ if $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. The limit does not exist at zero (see part (j)).

(d) False. Same counterexample as (c).

(e) True by Theorem 11.7.

(f) True (assuming the set is closed and bounded in $\mathbb{R}^n$), because a closed and bounded set in $\mathbb{R}^n$ is compact by the Heine-Borel theorem, and the continuous image of a compact set is compact by Theorem 10.33, which implies that it is closed.

(g) True. For a set E, we know that $\partial(E) = \overline{E} \setminus E^\circ$. Since $\overline{E}$ is closed and E° is open, we know that $\overline{E} \setminus E^\circ = \overline{E} \cap (\mathbb{R}^n \setminus E^\circ)$ is an intersection of closed sets and is closed.

(h) True by Theorem 10.17.

(i) True. A function $f : V \to \mathbb{R}^m$, where V is open in $\mathbb{R}^n$ is differentiable at a point $\mathbf{x}$ if and only if $\lim_{\mathbf{h} \to \mathbf{0}} \dfrac{|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - Df(\mathbf{x})\mathbf{h}|}{|\mathbf{h}|} = 0$, which is true if and only if $\lim_{\mathbf{h} \to \mathbf{0}} \dfrac{f_i(\mathbf{x} + \mathbf{h}) - f_i(\mathbf{x}) - Df_i(\mathbf{x}) \cdot \mathbf{h}}{|\mathbf{h}|} = \mathbf{0}$ for each i, where $Df_i(\mathbf{x})$ represents the ith row of $Df(\mathbf{x})$ by Theorem 10.17, which is true if and only if f_i is differentiable for each component f_i of f.

(j) False. A counterexample would be $f(x, y) = \dfrac{2x^2y}{x^4 + y^2}$. Approaching the origin along any line would have a limit of zero, but approaching along $y = x^2$ would give a limit of 1.

(k) False. Unless it is known that at least one of f_{xy} or f_{yx} is continuous at the point in question, Clairaut's Theorem fails.

(l) False. The continuous image of a connected set is connected but not necessarily closed (just take the identity map $f(x) = x$, and note that $f((0, 1)) = (0, 1)$ is not closed).

(m) False. Let $f(x) = x^2$. Then $f^{-1}([1, 4]) = [-2, -1] \cup [1, 2]$, which is not connected.

(n) True. If the graph G of a function $f : [a, b] \to \mathbb{R}$ is closed and connected we know that then the function is continuous by Theorem 10.48, which means that $F(x) = (x, f(x))$ is also continuous on $[a, b]$ by Theorem 10.17. Hence $F([a, b]) = G$ is compact by Theorem 10.33.

(o) True. The union of two closed sets is closed, and the union of two bounded sets is bounded, so the union of two closed and bounded sets is closed and bounded. In $\mathbb{R}^n$ this means that the union is compact by the Heine-Borel Theorem.

(p) False. The graphs of $y = x^2$ and $y = 2 - x^2$ are connected, but their intersection is the disconnected set consisting of the points $(-1, 1)$ and $(1, 1)$.

(q) True by Theorem 10.30.

(r) True because the inverse of complement of a set is the complement of the inverse, so this follows from Theorem 10.30.

(s) False. Using $f(x) = 1$ is continuous on $\mathbb{R}$ and the single point set $\{1\}$ is compact, but $f^{-1}(\{1\}) = \mathbb{R}$, which is not compact.

□

Chapter 12

Integration in Higher Dimensions

When integrating a function over a domain higher dimensional Euclidean spaces, we have to address the question of which functions can be integrated and the question of which domains may be integrated over, though this can be reduced to considering which functions may be integrated over n-rectangles since integrating over a subset of such a rectangle can simply be thought of as integrating a function which is re-defined to be zero outside of that subset. Thus, we begin with integrating over a rectangle in $\mathbb{R}^n$.

We begin with a lot of definitions and background theorems. A reader might get bogged down in this (necessarily extensive) foundation that we require to prove the main theorems we want to use about integration and unify ideas discussed earlier. For instance, we will have notions about volume based on integration over a region and notions based on outer and inner sums and we need to demonstrate that these are the same or our concept of volume is inconsistent.

Definition 86

If g is a real valued function whose domain includes $\{1, 2, 3, ..., k\}$ for some $k \in \mathbb{N}$ then we use the notation $\prod_{i=1}^{k} g(i) = g(1)g(2)...g(k)$. If $S_1, S_2, ..., S_k$ are sets then we use $\prod_{i=1}^{k} S_i$ to denote the Cartesian product $S_1 \times S_2 \times ... \times S_k$ whose elements are k-tuples $(s_1, s_2, s_3, ..., s_k)$ where each $s_i \in S_i$ for $1 \leq i \leq k$.

We define an *n-rectangle* (or just a rectangle) to be the Cartesian product of n closed intervals $R = \prod_{i=1}^{n} [a_i, b_i]$. We say the *$n$-volume* (or just the volume) of R is $|R| = \prod_{i=1}^{n} b_i - a_i$. We may refer to $[a_i, b_i]$ as the ith *edge factor* of R. An *n-cube* or *cube* in $\mathbb{R}^n$ is a rectangle whose edge factors all have equal length. We will refer to $C_\epsilon(\mathbf{x}) = \prod_{i=1}^{n} (x_i - \frac{\epsilon}{2}, x_i + \frac{\epsilon}{2})$ as the ϵ-cube centered at $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$.

A rectangle in $\mathbb{R}^2$ is the normal idea of a rectangle plus the region enclosed by that rectangle (also called a rectangular disk). A rectangle in $\mathbb{R}$ is just a line segment, and a rectangle in $\mathbb{R}^3$ is a box plus the region enclosed by the box. Similarly, a cube in $\mathbb{R}^2$ is a square plus the region enclosed by the square. When integrating a function of multiple variables we multiply function values by the volumes of rectangles rather than the length of line segments in the domain as we did for single variable functions.

Definition 87

Let $f : R \to \mathbb{R}$, where $R = \prod_{i=1}^{n}[a_i, b_i]$ is a rectangle in $\mathbb{R}^n$. For partitions $P_k = \{x_0^{(k)}, x_1^{(k)}, ..., x_{n_k}^{(k)}\}$ of $[a_k, b_k]$ for $1 \leq k \leq n$, we refer to $P = \{P_1, P_2, P_3, ..., P_n\}$ as a *partition* of R, and we refer to $G = G(P) = \{\prod_{j=1}^{n}[x_{i_j-1}^{(j)}, x_{i_j}^{(j)}] | i_j \in \{1, 2, ..., n_j\}$ for each $1 \leq j \leq n\}$ as the *grid* on R (or over R) induced by P. The *mesh* $|G|$ of grid G is the largest diameter of any element of G.

Let partitions $Q_1, Q_2, ..., Q_n$ be partitions of $[a_1, b_1], [a_2, b_2], ..., [a_n, b_n]$ which induce grid H on $R = \prod_{i=1}^{n}[a_i, b_i]$. Let partitions $P_1, P_2, ..., P_n$ be partitions of $[a_1, b_1], [a_2, b_2], ..., [a_n, b_n]$ which induce grid G on R. If partitions $Q_1, Q_2, ..., Q_n$ are refinements of partitions $P_1, P_2, P_3, ..., P_n$ respectively, then we say that H is a *refinement* of G and that partition $Q = \{Q_1, Q_2, Q_3, ..., Q_n\}$ is a refinement of partition $P = \{P_1, P_2, P_3, ..., P_n\}$. We will use the notation $G * H$ to denote the refinement of G and H induced by partition $P * Q = \{P_1 \cup Q_1, P_2 \cup Q_2, ..., P_n \cup Q_n\}$.

For each $R_t \in G$ we let $M_t = \sup_{\mathbf{x} \in R_t} f(\mathbf{x})$ and $m_t = \inf_{\mathbf{x} \in R_t} f(\mathbf{x})$. We define the *upper sum* of f with respect to grid G to be $U(f, G) = \sum_{R_t \in G} M_t |R_t|$, and the *lower sum* of f with respect to grid G to be $L(f, G) = \sum_{R_t \in G} m_t |R_t|$.

Here, the idea of a grid takes the place of the notion of a partition of a closed interval in one dimension.

Definition 88

We refer to the *upper integral* of f to be $(U)\int_R f = \inf_G U(f, G)$ (where G is understood to range over all possible grids on R). We refer to the *lower integral* of f to be $(L)\int_R f = \sup_G L(f, G)$. If $(L)\int_R f = (U)\int_R f = I$ then we say that f is *integrable* and the *integral* of f on R (or over R) is $\displaystyle\int_R f = I$.

If $T_k = \{x_1^{(k)^*}, x_2^{(k)^*}, ..., x_{n_k}^{(k)^*}\}$ is a marking of P_k for each $1 \leq k \leq n$ then we will refer to $T = T(P) = \displaystyle\prod_{i=1}^{n} T_i$ as a *marking* of the grid G. We may shorten notation by saying that if $R_t = \displaystyle\prod_{j=1}^{n}[x_{i_j-1}^{(j)}, x_{i_j}^{(j)}] \in G$ then $\mathbf{x}_t^* = (x_{i_1}^{(1)^*}, x_{i_2}^{(2)^*}, ..., x_{i_n}^{(n)^*}) \in T$. We say that $S_T(f, G) = \displaystyle\sum_{R_t \in G} f(\mathbf{x}_t^*)|R_t|$ is the *Riemann sum* of f over G with respect to marking T.

Note that, for each $R_t \in G$, it is always true that if $\mathbf{x}, \mathbf{y} \in R_t$ then $|\mathbf{x} - \mathbf{y}| \leq |G|$.

We tend to use iterated integral signs to refer to the integral of an integral, but iterated integrals signs without bounds are also used simply to declare the dimension of the domain to be integrated over. Thus, if R is a two dimensional rectangle we may write $\displaystyle\int_R f = \int\int_R f = \int\int_R f dA = \int\int_R f(x, y)dA$. All of these mean the same thing, and while the "dA" is not a required convention, using the letter A is intended to suggest to the reader that the grid rectangles have a two dimensional volume (an area). Likewise, if R is a 3-rectangle then we use a variety of equivalent notations $\displaystyle\int_R f = \int\int\int_R f = \int\int\int_R f dV = \int\int\int_R f(x, y, z)dV$, where the letter V is intended to make the reader think of of the grid rectangles as having a three (or higher) dimensional volume.

In dimension two these notions are easiest to picture. A rectangle $R = [a_1, b_1] \times [a_2, b_2]$, which is the set of all points $(x, y) \in \mathbb{R}^2$ so that $a_1 \leq x \leq b_1$ and $a_2 \leq y \leq b_2$. Readers are encouraged to take a few minutes thinking about what each of these definitions would look like for a two dimensional rectangle (integrals of positive functions over these rectangles are volumes, so it is reasonable to see the definitions visually).

Our first theorem addresses ideas of containment and convexity. Each ball contains a cube, and each cube contains a ball about a point. Rectangles are convex, and so are balls. We can use open cubes as a basis for a topology on $\mathbb{R}^n$ instead of open balls when it is convenient.

Theorem 12.1. *Some basic properties of rectangles and distance:*

(a) Let $\boldsymbol{p} = (p_1, p_2, p_3, ..., p_n) \in \mathbb{R}^n$ *and let* $\epsilon > 0$. *Then the* ϵ*-cube* $C_\epsilon(\boldsymbol{p})$ *has diameter* $\epsilon\sqrt{n}$, *and* $C_\epsilon(\boldsymbol{p}) \subseteq B_{\epsilon\sqrt{n}}(\boldsymbol{p})$ *and* $B_\epsilon(\boldsymbol{p}) \subseteq C_{2\epsilon}(\boldsymbol{p})$.

(b) Let $R = \displaystyle\prod_{i=1}^{n}[a_i, b_i]$ *be an* n*-rectangle in* $\mathbb{R}^n$, *and let* $\boldsymbol{a} = (a_1, a_2, a_3, ..., a_n)$ *and* $\boldsymbol{b} = (b_1, b_2, b_3, ..., b_n)$. *Then the diameter of* R *is* $|\boldsymbol{a} - \boldsymbol{b}|$. *Furthermore,* R *is convex.*

(c) A set $U \subseteq \mathbb{R}^n$ is open if and only if for every $\boldsymbol{p} \in U$ there is and $\epsilon > 0$ so that $C_\epsilon(\boldsymbol{p}) \subseteq U$.

Proof. (a) Since $|(p_1 - \frac{\epsilon}{2}, p_2 - \frac{\epsilon}{2}, ..., p_n - \frac{\epsilon}{2}) - (p_1 + \frac{\epsilon}{2}, p_2 + \frac{\epsilon}{2}, ..., p_n + \frac{\epsilon}{2})| = \sqrt{\sum_{i=1}^{n} \epsilon^2} = \epsilon\sqrt{n}$, we know that $diam(\overline{C_\epsilon(\mathbf{p})}) = diam(C_\epsilon(\mathbf{p})) \geq \epsilon\sqrt{n}$. For any $\mathbf{x}, \mathbf{y} \in C_\epsilon(\mathbf{p})$ we know that $|x_i - y_i| < \epsilon$ for each $1 \leq i \leq n$, which means that $|\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} < \epsilon\sqrt{n}$ and therefore $diam(C_\epsilon(\mathbf{p})) = \epsilon\sqrt{n}$.

We note that for any $\mathbf{x} \in C_\epsilon(\mathbf{p})$, for each $1 \leq i \leq n$, we know $|x_i - p_i| < \frac{\epsilon}{2}$, so $|\mathbf{x} - \mathbf{p}| \leq \frac{\epsilon\sqrt{n}}{2}$ so $C_\epsilon(\mathbf{p}) \subseteq B_{\epsilon\sqrt{n}}(\mathbf{p})$.

For any $\mathbf{x} \in B_\epsilon(\mathbf{p})$ we know that $|x_i - p_i| \leq |\mathbf{x} - \mathbf{p}| < \epsilon$ for each $1 \leq i \leq n$, which means that $\mathbf{x} \in C_{2\epsilon}(\mathbf{p})$.

(b) We know that the diameter of R is at least $|\mathbf{a} - \mathbf{b}|$ since $\mathbf{a}, \mathbf{b} \in R$. For any $\mathbf{x} = (x_1, x_2, x_3, ..., x_n), \mathbf{y} = (y_1, y_2, y_3, ..., y_n) \in R$ we also know that $x_i, y_i \in [a_i, b_i]$ for each $1 \leq i \leq n$. Therefore, $|x_i - y_i| \leq |a_i - b_i|$ which means that $|\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \leq |\mathbf{a} - \mathbf{b}|$, so $diam(R) = |\mathbf{a} - \mathbf{b}|$ as desired.

To see that R is convex, let $\mathbf{x}, \mathbf{y} \in R$. Then $L(\mathbf{x}, \mathbf{y}) = \{(1-t)\mathbf{x} + t\mathbf{y} | 0 \leq t \leq 1\}$. For each $1 \leq i \leq n$ and each $0 \leq t \leq 1$ we note that if $x_i \leq y_i$ then $(1-t)x_i + ty_i = x_i + t(y_i - x_i) \geq x_i$ since $y_i - x_i \geq 0$ and also $(1-t)x_i + ty_i \leq (1-t)y_i + ty_i = y_i$ since $1 - t \geq 0$. Similarly, if $x_i \geq y_i$ then $(1-t)x_i + ty_i \in [y_i, x_i]$. Since $a_i \leq x_i$ and $a_i \leq y_i$, and $x_i \leq b_i$ and $y_i \leq b_i$ for each $1 \leq i \leq n$, it follows that $(1-t)x_i + ty_i \in [a_i, b_i]$ for each $1 \leq i \leq n$ and $0 \leq t \leq 1$ and therefore $L(\mathbf{x}, \mathbf{y}) \subseteq R$. Thus, R is convex.

(c) Let U be open and let $\mathbf{p} \in U$. Choose $\epsilon > 0$ so that $B_\epsilon(\mathbf{p}) \subseteq U$. Then by part (a) we know that $C_{\frac{\epsilon}{\sqrt{n}}} \subseteq B_\epsilon(\mathbf{p}) \subseteq U$.

Let U be a set so that every point in U is contained in an epsilon cube centered at that point and contained in U. Let $\mathbf{p} \in U$ and choose $\epsilon > 0$ so that $C_\epsilon(\mathbf{p}) \subseteq R$. Then by part (a) we know that $B_{\frac{\epsilon}{2}}(\mathbf{p}) \subseteq C_\epsilon(\mathbf{p}) \subseteq R$, so U is open.

□

Our next theorem addresses the usual ordering of sums for a grid, that lower sums are no more than Riemann sums, which are no more and upper sums over any particular grid.

Theorem 12.2. *Let $f : R \to \mathbb{R}$ be bounded, where R is a rectangle in $\mathbb{R}^n$. Let G be a grid over R. Let T be a marking of G. Then $L(f, G) \leq S_T(f, G) \leq U(f, G)$.*

Proof. For each $R_t \in G$ we know that $m_t \leq f(\mathbf{x}_t) \leq M_t$, from which it follows that $\sum_{R_t \in G} m_t|R_t| \leq \sum_{R_t \in G} f(\mathbf{x}_t)|R_t| \leq \sum_{R_t \in G} M_t|R_t|$.

□

As in one dimension, we have to establish that refining a grid will make upper sums decrease (or stay the same) and lower sums increase (or stay the same), which is the objective of the next theorem.

Theorem 12.3. *Let $f : R \to \mathbb{R}$ be bounded, where $R = \prod_{i=1}^{n}[a_i, b_i]$ is an n-rectangle in $\mathbb{R}^n$. Let $P = \{P_1, P_2, P_3, ..., P_n\}$, where $P_k = \{x_0^{(k)}, x_1^{(k)}, ..., x_{n_k}^{(k)}\}$ is a partition of $[a_k, b_k]$ for $1 \leq k \leq n$. Let $Q = \{Q_1, Q_2, Q_3, ..., Q_n\}$, where $Q_k = \{q_0^{(k)}, q_1^{(k)}, ..., q_{n_k}^{(k)}\}$ is a partition of $[a_k, b_k]$ for $1 \leq k \leq n$, and let $G = G(P)$ and $H = H(Q)$ be the grids induced by these partitions, where H is a refinement of G. Then $L(f, G) \leq L(f, H) \leq U(f, H) \leq U(f, G)$.*

Proof. We first prove the result is true when $Q_j = P_j \cup \{q\}$, where $q \in (x_{s-1}^{(j)}, x_s^{(j)})$, and $P_i = Q_i$ for $i \neq j$. Let S be the set of all elements of grid G whose jth edge factor is $[x_{s-1}^{(j)}, x_s^{(j)}]$. Let $R_t = [x_{i_1-1}^{(1)}, x_{i_1}^{(1)}] \times [x_{i_2-1}^{(2)}, x_{i_2}^{(2)}] \times ... \times [x_{s-1}^{(j)}, x_s^{(j)}] \times ... \times [x_{i_n-1}^{(n)}, x_{i_n}^{(n)}] \in S$. Then $R_t = R_t^{(L)} \cup R_t^{(U)}$, where $R_t^{(L)} = [x_{i_1-1}^{(1)}, x_{i_1}^{(1)}] \times [x_{i_2-1}^{(2)}, x_{i_2}^{(2)}] \times ... \times [x_{s-1}^{(j)}, q] \times ... \times [x_{i_n-1}^{(n)}, x_{i_n}^{(n)}] \in H$ and $R_t^{(U)} = [x_{i_1-1}^{(1)}, x_{i_1}^{(1)}] \times [x_{i_2-1}^{(2)}, x_{i_2}^{(2)}] \times ... \times [q, x_s^{(j)}] \times ... \times [x_{i_n-1}^{(n)}, x_{i_n}^{(n)}] \in H$.

If we let $M_t^{(L)} = \sup_{R_t^{(L)}} f(\mathbf{x})$ and $m_t^{(L)} = \inf_{R_t^{(L)}} f(\mathbf{x})$, and we let $M_t^{(U)} = \sup_{R_t^{(U)}} f(\mathbf{x})$ and $m_t^{(U)} = \inf_{R_t^{(U)}} f(\mathbf{x})$, then by Theorem 1.17 we know that $M_t \geq \max\{M_t^{(L)}, M_t^{(U)}\}$ and $m_t \leq \min\{m_t^{(L)}, m_t^{(U)}\}$ (where $M_t = \sup_{R_t} f(\mathbf{x})$ and $m_t = \inf_{R_t} f(\mathbf{x})$).

Thus, it follows that $U(f, G) - U(f, H) = \sum_{R_t \in S} M_t|R_t| - M_t^{(L)}|R_t^{(L)}| - M_t^{(U)}|R_t^{(U)}| \geq \sum_{R_t \in S} M_t(|R_t| - |R_t^{(L)}| - |R_t^{(U)}|) = 0$, so $U(f, G) \geq U(f, H)$. Likewise, $L(f, G) - L(f, H) = \sum_{R_t \in S} m_t|R_t| - m_t^{(L)}|R_t^{(L)}| - m_t^{(U)}|R_t^{(U)}| \leq \sum_{R_t \in S} m_t(|R_t| - |R_t^{(L)}| - |R_t^{(U)}|) = 0$, so $L(f, G) \leq U(f, H)$.

Next, let $P_i = Q_i$ for $i \neq j$ and let $Q_j = P_i \cup \{q_1, q_2, ..., q_m\}$. Then let H_i be the grid induced by $P_1, P_2, ..., P_{j-1}, P_j \cup \{q_1, q_2, ..., q_i\}, P_{j+1}, ..., P_n\}$ for all $1 \leq i \leq m$. Then we see that $L(f, G) \leq L(f, H_1) \leq L(f, H_2) \leq ... \leq L(f, H_m) = L(f, H) \leq U(f, H) = U(f, H_m) \leq U(f, H_{m-1}) \leq ... \leq U(f, G)$.

Finally, let $Q_s = P_s \cup D_s$ for some finite set D_s contained in $[a_s, b_s]$ for each $1 \leq s \leq n$. We define grid K_w, for each positive integer $1 \leq w \leq n$, to be the grid induced by $P_s \cup D_s$ for $1 \leq s \leq w$ and P_s for $w < s \leq n$. By the preceding paragraph, we have that $L(f, G) \leq L(f, K_1) \leq L(f, K_2) \leq ... \leq L(f, K_n) = L(f, H) \leq U(f, H) = U(f, K_n) \leq U(f, K_{n-1}) \leq ... \leq U(f, K_1) \leq U(f, G)$ as desired.

□

Paralleling the development in single variable calculus, we next establish that lower sums never exceed upper sums, even if they are upper and lower sums over different grids.

Theorem 12.4. *Let $f : R \to \mathbb{R}$, where $R = \prod_{i=1}^{n}[a_i, b_i]$ is an n-rectangle in $\mathbb{R}^n$, be bounded. Let G and H be grids on R. Then $L(f, G) \leq U(f, H)$. Furthermore, $(L)\int_R f \leq (U)\int_R f$.*

Proof. If P_i, Q_i are partitions on $[a_i, b_i]$ for $1 \leq i \leq n$ so $G = G(P_1, P_2, P_3, ..., P_n)$ and $H = H(Q_1, Q_2, Q_3, ..., Q_n)$, then $G * H$ is a grid which refines both G and H. By Theorem 12.3 it follows that $L(f, G) \leq L(f, G * H) \leq U(f, G * H) \leq U(f, H)$.

Thus, for any grid H on R, it is the case that $U(f, H) \geq L(f, G)$ for all grids G on R, so $U(f, H) \geq (L)\int_R f$, which makes $(L)\int_R f$ a lower bound for all upper sums $U(f, H)$, which means that $(L)\int_R f \leq (U)\int_R f$.

□

As with single variable integrals, we can next establish that a function is integrable if and only if the upper and lower sums can be made arbitrarily close to one another, which is the next theorem.

Theorem 12.5. *Let $f : R \to \mathbb{R}$ be bounded, where R is a rectangle in $\mathbb{R}^n$. Then f is integrable if and only if for every $\epsilon > 0$ there is a grid G on R so that $U(f, G) - L(f, G) < \epsilon$.*

Proof. First, assume that f is integrable. By the approximation property, we can find grids G, H so that $U(f, G) < (U)\int_R f + \epsilon$ and $L(f, H) > (U)\int_R f - \epsilon$, which means that $\int_R f - \epsilon < L(f, H) \leq L(f, G * H) \leq U(f, G * H) \leq U(f, G) < \int_R f + \epsilon$. Thus, it follows that $U(f, G * H) - L(f, G * H) < \epsilon$.

Next, assume that for every $\epsilon > 0$ there is a grid G on R so that $U(f, G) - L(f, G) < \epsilon$. Let $\epsilon > 0$. Choose G so that $U(f, G) - L(f, G) < \epsilon$. But then we know that $L(f, G) \leq (L)\int_R f \leq (U)\int_R f \leq U(f, G)$, so $0 \leq (U)\int_R f - (L)\int_R f < \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $(L)\int_R f = (U)\int_R f$, so f is integrable.

□

We next show that continuous functions on a rectangle are integrable, and that if a grid mesh is sufficiently small then upper and lower sums can be made arbitrarily close to one another.

Theorem 12.6. *Let $f : R \to \mathbb{R}$ be continuous, where R is a rectangle in $\mathbb{R}^n$. Then f is integrable. Furthermore, for any $\epsilon > 0$ we can find a number $\delta > 0$ so that if G is any grid over R with $|G| < \delta$ then $U(f, G) - L(f, G) < \epsilon$.*

Proof. Since f is continuous on the closed and bounded rectangle R, from Theorem 10.35 we know that f is uniformly continuous. Let $\epsilon > 0$. Choose $\delta > 0$ so that if $|\mathbf{x} - \mathbf{y}| < \delta$ then $|f(\mathbf{x}) - f(\mathbf{y})| < \frac{\epsilon}{|R|}$. Let G be a grid over R with $|G| < \delta$. By the Extreme Value Theorem there are points $\mathbf{p}_t, \mathbf{q}_t \in R_t$ for each rectangle $R_t \in G$ so that $f(\mathbf{q}_n) \leq f(\mathbf{x}) \leq f(\mathbf{p}_n)$ for each $\mathbf{x} \in R_t$. Hence, $U(f,G) - L(f,G) = \sum_{R_t \in G} (f(\mathbf{p}_t) - f(\mathbf{q}_t))|R_t| < \frac{\epsilon}{|R|} \sum_{R_t \in G} |R_t| = \epsilon$, which means that f is integrable. □

We would like to integrate over regions that are not rectangles. To do this, we just extend the function to be defined on a rectangle by defining the function to be zero elsewhere, but it would be nice to know which regions continuous functions can be restricted to and continue to be integrable over those regions. The ideas of Lebesgue measure zero and Jordan content will help us to understand which regions can be integrated over and which functions are integrable.

Definition 89

Let $S \subseteq \mathbb{R}^n$. We say that S has *Lebesgue measure zero*, written $\lambda(S) = 0$, if, for every $\epsilon > 0$ there is a countable collection of rectangles $\{R_i\}_{i \in \mathbb{N}}$ (the collection can also be finite) which covers S so that $\sum_{i=1}^{\infty} |R_i| < \epsilon$. We say that S has *Jordan content zero* or *volume zero*, denoted $Vol(S) = 0$, if, for every $\epsilon > 0$ there is a finite collection of rectangles $\{R_i\}_{1 \leq i \leq m}$ which covers S so that $\sum_{i=1}^{m} |R_i| < \epsilon$.

If $E \subseteq R$, an n-rectangle, and $G = \{R_t\}_{1 \leq t \leq k}$ is a grid on R then we define $O(E,G)$ to denote $\{R_j \in G | R_j \cap \overline{E} \neq \emptyset\}$, the set of *outer rectangles* for E, $I(E,G)$ to denote $\{R_j \in G | R_j \subseteq E^\circ\}$, the set of *inner rectangles*, and $S(E,G) = \{R_j \in G | R_j \cap E \neq \emptyset\}$ the set if *intersecting rectangles* for E with respect to grid G.

We define $V(E,G) = \sum_{R_t \in O(E,G)} |R_t|$ to be the *outer sum* of E with respect to the grid G, and $v(E,G) = \sum_{R_t \in I(E,G)} |R_t|$ to be the *inner sum* of E with respect to the grid G. We define the *outer volume* of E to be $(O)Vol(E) = \inf_G V(E,G)$ and the *inner volume* of E to be $(I)Vol(E) = \sup_G v(E,G)$.

Sets S and T are *non-overlapping* if $Vol(S \cap T) = 0$. If it is false that S and T are non-overlapping then we say that S and T *overlap*.

Note that in the above definition a finite union of rectangles is always closed, which means that any collection of rectangles covering E also covers $\overline{E}$. Hence, if we let $\mathcal{D}$ be the set of all finite sets of rectangles which cover E. Then $(O)Vol(E) = \inf_{\mathcal{C} \in \mathcal{D}} \sum_{R \in \mathcal{C}} |R|$ (because $\mathcal{D}$ is also the collection of all finite sets of rectangles that cover $\overline{E}$). Thus, $(O)Vol(E) = 0$ if and only if $Vol(E) = 0$.

Also, in the definition of rectangle above, we have only defined a set to be an n-rectangle if it is the product of edge factors in the axes. However, by re-orienting the axes we can extend this definition to a product relative to any orthogonal set of coordinate directions. The terms “inner rectangles” and ‘outer rectangles” will be useful for us, but they are not standard terms used in other texts, and the choice $S(E,G)$ is supposed to make the reader think of rectangles "sharing" points with E.

In terms of these sets, $U(f,G)=\sum_{R_j\in S(E,G)} M_j|R_j|$ and $L(f,G)=\sum_{R_j\in S(E,G)} m_j|R_j|$.

At this point we also have two notions of volume for a rectangle (product of side lengths or infimum of sums of volumes of rectangles in finite covers of the rectangle). One of the things we will have to resolve is to show that these two definitions are equivalent.

First, it is helpful to notice that while there are some topological advantages to an open or closed cover by rectangles or interiors of rectangles, these ideas could be used interchangeably for establishing volume or measure zero, which is the objective of the next three theorems.

We first demonstrate that we can fatten a rectangle slightly to create a rectangle whose interior contains the first rectangle, increasing the volume by an arbitrarily small amount. In some situations it can be advantageous if the larger rectangle has rational side length (because such a rectangle can be subdivided into smaller cubes of equal side length).

Theorem 12.7. *Let $R=\prod_{i=1}^{n}[a_i,b_i]$ be a rectangle in $\mathbb{R}^n$ and let $\epsilon>0$. Then there is a number l so that if $0<d_i\le l$ and $0<c_i\le l$ for each $1\le i\le n$ then $Q=\prod_{i=1}^{n}[a_i-c_i,b_i+d_i]$ is a rectangle so that $R\subset Q^\circ$ and $Vol(Q)-Vol(R)<\epsilon$. If R is a cube then we can choose c_i and d_i so that Q is also a cube centered at the same point as R. We can also choose the c_i and d_i so that each a_i-c_i and each b_i+d_i is rational.*

Proof. First, note that $g:\mathbb{R}^n\to\mathbb{R}$ defined by $g(x_1,x_2,...,x_n)=x_1x_2x_3...x_n$ is a product of continuous functions and is therefore a continuous function, and in particular g is continuous at $(\mathbf{b}-\mathbf{a})$, where $\mathbf{b}=(b_1,b_2,b_3,...,b_n)$ and $\mathbf{a}=(a_1,a_2,a_3,...,a_n)$. Hence, we can find $\delta>0$ so that if $|\mathbf{x}-(\mathbf{b}-\mathbf{a})|<\delta$ then $|g(\mathbf{x})-g(\mathbf{a})|<\epsilon$. In particular, if we set $\mathbf{c}=(b_1-a_1+\frac{\delta}{2\sqrt{n}},b_2-a_2+\frac{\delta}{2\sqrt{n}},...,b_n-a_n+\frac{\delta}{2\sqrt{n}})$ then $|\mathbf{b}-\mathbf{a}-\mathbf{c}|=\sqrt{\sum_{i=1}^{n}\frac{\delta^2}{2^n n}}<\delta$, so $g(\mathbf{c})-g(\mathbf{b}-\mathbf{a})<\epsilon$. Setting $Q=\prod_{i=1}^{n}[a_i-\frac{\delta}{4\sqrt{n}},b_i+\frac{\delta}{4\sqrt{n}}]$ we have that $Vol(Q)=g(\mathbf{c})$ and $Vol(R)=g(\mathbf{b}-\mathbf{a})$, so $Vol(Q)-Vol(R)<\epsilon$. Notice that if we replace Q by $\prod_{i=1}^{n}[a_i-d_i,b_i+d_i]$ where $0<d_i\le\frac{\delta}{2\sqrt{n}}$ then each edge factor is no larger than before, so $Vol(R)<Vol(Q)<Vol(R)+\epsilon$. Also, if R is a cube and each $d_i=c_i=d<l$ then the new rectangle $Q=\prod_{i=1}^{n}[a_i-d,b_i+d]$ is a cube (since each side length was increased by the same amount).

Let $m=\min\{d_1,d_2,...,d_n\}$. For any $\mathbf{x}=(x_1,x_2,x_3,...,x_n)\in R$ we note that $B_m(\mathbf{x})\subset\prod_{i=1}^{n}(x_i-m,x_i+m)\subset Q$, so $R\subset Q^\circ$.

Finally, if we choose q_i, a rational number between b_i and b_i+l then we can set $d_i=q_i-b_i<1$ so that $b_i+d_i=q_i$, a rational number. Likewise, we can choose c_i so that $a-c_i$ is rational if we wish.

□

Since we can fatten rectangles and cubes slightly, we can also increase the volumes of covers consisting of such rectangles and cubes slightly, placing rectangles in the original cover into the interiors of rectangles in the new cover.

Theorem 12.8. *Let $E \subseteq \mathbb{R}^n$ and let $\epsilon > 0$. If there is a countable collection of rectangles $\{R_i\}_{i\in\mathbb{N}}$ which covers E so that $\sum_{i=1}^{\infty} |R_i| < \epsilon$ then there is a countable collection of rectangles $\{Q_i\}_{i\in\mathbb{N}}$ so that $R_i \subset Q_i^\circ$ for each $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty} |Q_i| < \epsilon$.*

Likewise, if there is a finite collection of rectangles $\{R_i\}_{1\leq i\leq k}$ which covers E so that $\sum_{i=1}^{k} |R_i| < \epsilon$ then there is a finite collection of rectangles $\{Q_i\}_{1\leq i\leq k}$ so that $R_i \subset Q_i^\circ$ for each $i \in \{1, 2, 3, ..., k\}$ and $\sum_{i=1}^{k} |Q_i| < \epsilon$.

Proof. Let $\{R_i\}_{i\in\mathbb{N}}$ be a cover of E so that $\sum_{i=1}^{\infty} |R_i| = \gamma < \epsilon$. By Theorem 12.7, we can choose Q_i for each $i \in \mathbb{N}$ so that $R_i \subset Q_i$ and $Vol(Q_i) - Vol(R_i) < \frac{\epsilon - \gamma}{2^{i+1}}$. Hence, $\{Q_i^\circ\}_{i\in\mathbb{N}}$ covers E and $\sum_{i=1}^{\infty} |Q_i| \leq \sum_{i=1}^{\infty} |R_i| + \sum_{i=1}^{\infty} \frac{\epsilon - \gamma}{2^{i+1}} = \gamma + \frac{\epsilon - \gamma}{2} < \epsilon$.

The finite case is similar. Let $\{R_i\}_{1\leq i\leq k}$ be a finite cover of E by n-rectangles with $\sum_{i=1}^{k} |R_i| = \gamma < \epsilon$. By Theorem 12.7, we can choose n-rectangles $\{Q_i\}_{1\leq i\leq k}$ so that $R_i \subset Q_i^\circ$ and $Vol(Q_i) - Vol(R_i) < \frac{\epsilon - \gamma}{2^{i+1}}$. As before, $\sum_{i=1}^{k} |Q_i| < \epsilon$.

□

In the following theorem we notice that it would make no difference whether we used rectangles or open rectangles (interiors of rectangles) in the definitions of measure or volume.

Theorem 12.9. *Let $S \subseteq \mathbb{R}^n$.*

(a) $\lambda(S) = 0$ if and only if, for every $\epsilon > 0$, there is a countable collection of rectangle interiors $\{R_i^\circ\}_{i\in\mathbb{N}}$ which covers S so that $\sum_{i=1}^{\infty} |R_i| < \epsilon$.

(b) $Vol(S) = 0$ if, for every $\epsilon > 0$, there is a finite collection of rectangle interiors $\{R_i^\circ\}_{1\leq i\leq m}$ which covers S so that $\sum_{i=1}^{m} |R_i| < \epsilon$.

Proof. If, for every $\epsilon > 0$, there is a countable collection of rectangle interiors $\{R_i^\circ\}_{i\in\mathbb{N}}$ which covers S so that $\sum_{i=1}^{\infty} |R_i| < \epsilon$ then $\{R_i\}_{i\in\mathbb{N}}$ covers S, so $\lambda(S) = 0$. Likewise, if, for every $\epsilon > 0$ there is a finite collection of rectangle interiors $\{R_i^\circ\}_{1\leq i\leq m}$ which covers S so that $\sum_{i=1}^{m} |R_i| < \epsilon$, then $\{R_i\}_{1\leq i\leq m}$ covers S, so $Vol(S) = 0$.

Conversely, assume that $\{R_i\}_{i\in\mathbb{N}}$ is a countable cover of S by rectangles R_i so that $\sum_{i=1}^{\infty} |R_i| < \epsilon$. By Theorem 12.8, we know that there is a countable collection of rectangles $\{Q_i\}_{i\in\mathbb{N}}$ so that $R_i \subset Q_i^\circ$ for each $i \in \mathbb{N}$ and $\sum_{i=1}^{\infty} |Q_i| < \epsilon$.

If $\{R_i\}_{1\leq i\leq k}$ is a finite cover of S by rectangles R_i so that $\sum_{i=1}^{k} |R_i| < \epsilon$ then, again, by Theorem 12.8, we can find $\{Q_i\}_{1\leq i\leq k}$ so that $R_i \subset Q_i^\circ$ for each $i \in \{1, 2, 3, ..., k\}$ and $\sum_{i=1}^{k} |Q_i| < \epsilon$.

□

Definition 90

Let $E \subset \mathbb{R}^n$ be bounded. If $Vol(\partial(E)) = 0$ then we say that E is a *Jordan region*. If E is a Jordan region then we define the *volume* or *Jordan content* of E to be $Vol(E) = (O)Vol(E)$.

Note that in the above definition a finite union of rectangles is always closed, which means that any collection of rectangles covering E also covers $\overline{E}$. Hence, if we let $\mathcal{D}$ be the set of all finite sets of rectangles which cover E. Then $(O)Vol(E) = \inf_{\mathcal{C}\in\mathcal{D}} \sum_{R\in\mathcal{C}} |R|$ (because $\mathcal{D}$ is also the collection of all finite sets of rectangles that cover $\overline{E}$).

We will show later that a bounded set E is a Jordan region if and only if $(O)Vol(E) = (I)Vol(E)$, which means that we could have used inner or outer volume in this definition of volume, and volume is defined if and only if it is the same as both inner and outer volume.

It may be instructive to address why we care about Jordan regions for a moment. A Jordan region is a region on which a characteristic function (a function whose value is one on the Jordan region and zero elsewhere) is integrable. Jordan regions are thus regions on which all functions which are otherwise integrable on a rectangle containing those Jordan regions are always integrable on the Jordan region (meaning that the characteristic function on the Jordan region times the original function is still integrable on the rectangle). The Jordan regions, therefore, are the nice domains over which it is reasonable to restrict an integrable function's domain and still talk about the integral of the function over that domain. This will be developed more formally in the theorems that follow, but first we have to address ideas related to volume to formalize notions that most likely seem intuitive to us already.

While volume zero always implies measure zero, the converse is false (consider the rational numbers in the real line, for instance). However, for a compact set (like a rectangle) the two ideas are equivalent, as shown below.

Theorem 12.10. *Let $E \subset \mathbb{R}^n$. If $Vol(E) = 0$ then $\lambda(E) = 0$. If E is compact and $\lambda(E) = 0$ then $Vol(E) = 0$. A set W is a Jordan region if and only if W is bounded and $\lambda(\partial(W)) = 0$.*

Proof. Let $\epsilon > 0$. If $Vol(E) = 0$ then we can find a finite cover of E by rectangles, the sum of whose volumes is less than ϵ. Since this finite set is also countable, it follows that $\lambda(E) = 0$.

Let E be compact with $\lambda(E) = 0$ and let $\epsilon > 0$. Then by Theorem 12.9, there is a collection of rectangles $\{R_i\}_{i \in \mathbb{N}}$ so that $\{R_i^\circ\}_{i \in \mathbb{N}}$ covers E and $\displaystyle\sum_{i=1}^{\infty} |R_i| < \epsilon$. Since E is compact, there is a finite subcover $\{R_{n_1}^\circ, R_{n_2}^\circ, ..., R_{n_k}^\circ\}$ which covers E, so $\displaystyle\sum_{i=1}^{k} |R_{n_i}| < \epsilon$. Thus, $Vol(E) = 0$.

To see that a Jordan region W is bounded we note that it must be possible to cover W with a finite number of rectangles, the union of which must be bounded. For any bounded set W, since $\partial(W) = \overline{W} \setminus W^\circ$ is closed and bounded, it must follow that $\partial(W)$ is compact and therefore has volume zero if and only if it has Lebesgue measure zero. From this it follows that a set W is a Jordan region if and only if it is bounded and its boundary has measure zero.

□

Next we show that a subset of a measure zero (or volume zero) set is always measure zero (or volume zero respectively).

Theorem 12.11. *Let $A \subset B$ and let $\lambda(B) = 0$. Then $\lambda(A) = 0$. Likewise, if $Vol(B) = 0$ then $Vol(A) = 0$.*

Proof. Let $\epsilon > 0$. Assume $\lambda(B) = 0$. Then there is an open cover of B by a countable collection of rectangles $\{R_n\}$ so that $\displaystyle\sum_{n=1}^{\infty} |R_n| < \epsilon$. Since $\{R_n\}$ is also a cover for A it follows that $\lambda(A) = 0$. Replacing the countable collection of rectangles by a finite collection of rectangles, we see by a similar argument that if $Vol(B) = 0$ then $Vol(A) = 0$. □

We next demonstrate that countable sets always have measure zero (they do not always have volume zero).

Theorem 12.12. *Let $E = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, ...\}$ be countable. Then $\lambda(E) = 0$.*

Proof. Let R_i be a rectangle containing $\mathbf{p}_i$ of volume less than $\frac{\epsilon}{2^{i+1}}$. Then $\{R_i\}$ covers E and $\sum_{i=1}^{\infty} |R_i| < \epsilon$, so $\lambda(E) = 0$. □

Next, we show that a union of countably many sets of measure zero has measure zero.

Theorem 12.13. *Let $\lambda(E_i) = 0$ for each $i \in \mathbb{N}$. Then $\lambda(\bigcup_{i=1}^{\infty} E_i) = 0$.*

Proof. For each $i \in \mathbb{N}$ choose rectangles $\{R_{(i,n)}\}_{n\in\mathbb{N}}$ which cover E_i so that $\sum_{n=1}^{\infty} |R_{(i,n)}| < \frac{\epsilon}{2^{i+2}}$. Then $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |R_{(i,j)}| < \epsilon$ and $\{R_{(i,j)}\}_{i,j\in\mathbb{N}}$ is a cover for $\bigcup_{i=1}^{\infty} E_i$, which has Lebesgue measure zero. □

We next note that the closure of a Jordan region is always a Jordan region of equal volume, and that a set of outer volume zero is always a Jordan region (of volume zero).

Theorem 12.14. *Let E be a Jordan region in $\mathbb{R}^n$. Then $\overline{E}$ is a Jordan region and $Vol(E) = Vol(\overline{E})$. Furthermore, if $(O)Vol(E) = 0$ then E is a Jordan region of volume zero.*

Proof. Since $\partial(E) = \overline{E} \setminus E^\circ$ and $\partial(\overline{E}) = \overline{E} \setminus \overline{E}^\circ$ and $E^\circ \subseteq \overline{E}^\circ$, it follows that $\partial(\overline{E}) \subseteq \partial(E)$. Since E is a Jordan region, $Vol(\partial(E)) = 0$, which means that $Vol(\partial(\overline{E}) = 0$ by Theorem 12.11, which means that $\overline{E}$ is a Jordan region.

Let $\mathcal{D}$ be any finite collection of rectangles that covers E. Then since $\bigcup \mathcal{D}$ is closed and contains E, we know $\bigcup \mathcal{D}$ also contains $\overline{E}$. Likewise, any finite collection of rectangles that covers $\overline{E}$ also covers E. Since the set of finite collections of rectangles covering E and set of the finite collections of rectangles that cover $\overline{E}$ are the same, $Vol(E) = Vol(\overline{E})$.

If $(O)Vol(E) = 0$ then $Vol(\overline{E}) = 0$, so $Vol(\partial(E)) = 0 = Vol(E)$ by Theorem 12.11. □

It is a consequence of the fact that rectangles are connected that if a rectangle intersects a set S and is not contained in the interior of S then the rectangle intersects the boundary of S, as described below.

Theorem 12.15. *Let $E \cap R \neq \emptyset$, where R is a rectangle in $\mathbb{R}^n$. Then R is connected, and $R \cap \partial(E) \neq \emptyset$ if and only if $R \nsubseteq E^\circ$.*

Proof. We know R is convex by Theorem 12.1 and therefore connected by Theorem 10.46.

Assume that $R \nsubseteq E^\circ$ and $R \cap E \neq \emptyset$. Suppose $R \cap \partial(E) = \emptyset$. Then $R \cap E^\circ \neq \emptyset$.

Let $H = R \cap E^\circ$ and let $K = R \setminus H$. Note that H and K are disjoint, non-empty and their union is R. Also, K contains no points of E and thus no limit points of H since there are no boundary points of E in K.

If $\mathbf{p} \in H$ then there is some $\delta > 0$ so that $B_\delta(\mathbf{p}) \subseteq E$ since $\mathbf{p} \in E^\circ$, which means that $\mathbf{p}$ is not a limit point of K since K contains no points of E. Thus, H and K separate R, which is impossible since R is connected. We conclude that R must intersect the boundary of E.

We know that if $R \subseteq E^\circ$ then $R \cap \partial(E) = 0$ by definition of boundary. □

We now show that if one Jordan region is contained in a second one then the volume of the superset is at least as large as the volume of the subset.

Theorem 12.16. *Let W and E be Jordan regions and let $W \subseteq E$. Then $Vol(W) \leq Vol(E)$.*

Proof. Let $\mathcal{D}_E$ be the set of finite collections of rectangles that cover E and let $V_E = \{\sum\limits_{R\in\mathcal{C}} |R||\mathcal{C} \in \mathcal{D}_E\}$, the set of all sums of volumes of elements of $\mathcal{D}_E$. Let $\mathcal{D}_W$ be the set of finite collections of rectangles that cover W and let $V_W = \{\sum\limits_{R\in\mathcal{C}} |R||\mathcal{C} \in \mathcal{D}_E\}$, the set of all sums of volumes of elements of $\mathcal{D}_W$.

Any cover of E is also a cover of W, which means that $V_E \subseteq V_W$. Hence, $Vol(W) = \inf(V_W) \leq \inf(V_E) = Vol(E)$.

□

Next, we explain why two rectangles (defined in the usual way with edge factors in the axes) will always have an intersection which is another rectangle if they overlap.

Theorem 12.17. *Let $R = \prod\limits_{i=1}^{n}[a_i, b_i]$ and $Q = \prod\limits_{i=1}^{n}[c_i, d_i]$ be rectangles whose interiors intersect in $\mathbb{R}^n$. Then $R \cap Q$ is a rectangle.*

Proof. By definition, $R \cap Q = \prod\limits_{i=1}^{n}[\max\{a_i, c_i\}, \min\{b_i, d_i\}]$. Since the intersection of the interiors is not empty, this is a rectangle. □

We now demonstrate that rectangles are Jordan regions.

Theorem 12.18. *Let $R = \prod\limits_{i=1}^{n}[a_i, b_i]$ be a rectangle in R^n. Then R is a Jordan region.*

Proof. Let $\mathbf{x} = (x_1, x_2, x_3, ..., x_n) \in R$. If $a_i < x_i < b_i$ for each $i \in \{1, 2, 3, ..., n\}$ then let $m = \min\{x_1 - a_1, b_1 - x_1, x_2 - a_2, b_2 - x_2, ..., x_n - a_n, b_n - x_n\}$. Then $B_m(\mathbf{x}) \subset \prod\limits_{i=1}^{n}(x_i - m, x_i + m) \subseteq R$, which means that $\mathbf{x} \in R^\circ$.

It follows that if $\mathbf{x} \in \partial(R)$ then $x_i = a_i$ or $x_i = b_i$ for some i. We define rectangles $R_1(\delta), R_2(\delta), ..., R_{2n}(\delta)$ by $R_{2j-1}(\delta) = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_{j-1}, b_{j-1}] \times [a_j - \frac{\delta}{2}, a_j +$

$\frac{\delta}{2}] \times [a_{j+1}, b_{j+1}] \times ... \times [a_n, b_n]$ and $R_{2j} = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_{j-1}, b_{j-1}] \times [b_j - \frac{\delta}{2}, b_j + \frac{\delta}{2}] \times [a_{j+1}, b_{j+1}] \times ... \times [a_n, b_n]$ for each $j \in \{1, 2, ..., n\}$. Then $\partial(E) \subset \bigcup_{i=1}^{2n} R_i$. Let $M = \prod_{i=1}^{n}(b_i - a_i + 1)$. Then $\sum_{i=1}^{2n} |R_i| < \sum_{i=1}^{2n} M\delta$. Hence, if we choose $\delta < \frac{\epsilon}{2nM}$ then $\sum_{i=1}^{2k} |R_i| < \epsilon$. Thus, $Vol(\partial(R)) = 0$.

□

We next prove that if we add the rectangle volumes in the grid rectangles induced by a grid then the sum of the volumes is the volume of the original rectangle on which the grid was taken.

Theorem 12.19. *Let $R = \prod_{i=1}^{n}[a_i, b_i]$ be a rectangle in $\mathbb{R}^n$ and let $G = G(P)$ be a grid on R, where $P = \{P_1, P_2, ..., P_n\}$ and $P_i = \{x_0^{(i)}, x_1^{(i)}, ..., x_{n_i}^{(i)}\}$ for each $1 \leq i \leq n$. Then $|R| = \sum_{R_t \in G} |R_t|$.*

Proof. We will induct on the dimension of R. If $n = 1$ then $R = [a_1, b_1]$, and G is just a partition $P_1 = \{x_0^{(1)}, x_1^{(1)}, ..., x_{n_1}^{(1)}\}$, and $\sum_{i=1}^{n_1} x_i^{(1)} - x_{i-1}^{(1)} = b_1 - a_1 = |R|$.

Next, assume that for some $k \in \mathbb{N}$ it is true that for any grid H on $\prod_{i=1}^{k}[a_i, b_i]$, the sum $\sum_{R_t \in H} |R_t| = \prod_{i=1}^{k}(b_i - a_i)$. Let $R = \prod_{i=1}^{k+1}[a_i, b_i]$ with grid $G = G(P_1, P_2, ..., P_k, P_{k+1})$ on R, where $H = H(P_1, P_2, ..., P_k)$. Then $G = \{R_s \times [x_{i-1}^{(k+1)}, x_i^{(k+1)}] | R_s \in H \text{ and } 1 \leq k \leq n_{k+1}\}$. Hence, $\sum_{R_t \in G} |R_t| = \sum_{i=1}^{n_{k+1}} \sum_{R_s \in H} |R_s|(x_i^{(k+1)} - x_{i-1}^{(k+1)}) = \sum_{i=1}^{n_{k+1}} (x_i^{(k+1)} - x_{i-1}^{(k+1)}) \prod_{i=1}^{k}(b_i - a_i) = \prod_{i=1}^{k+1}(b_i - a_i) = |R|$.

□

In the next theorem, we show that the sum of the volumes of rectangles in a grid on one rectangle which are contained in a second rectangle is less than or equal to the volume of the second rectangle.

Theorem 12.20. *Let $G = \{Q_t\}_{1 \leq t \leq k}$ is a grid on a rectangle W in $\mathbb{R}^n$. Let R be a rectangle in $\mathbb{R}^n$. Let $C = \{Q_t \in G | Q_t \subseteq R\}$. Then $\sum_{Q_t \in C} |Q_t| \leq |R|$.*

Proof. If R and W do not overlap then $C = \emptyset$ and the result follows. Assume that R and W overlap. Let $G = G(P)$ be a grid on W where $P = \{P_1, P_2, ..., P_n\}$ and $P_i =$

$\{x_0^{(i)}, x_1^{(i)}, ..., x_{n_i}^{(i)}\}$ for each $1 \leq i \leq n$. Let $R = \prod_{i=1}^{n} [a_i, b_i]$ for some a_i, b_i for $1 \leq i \leq n$. Then $H = H(P_1 \cap [a_1, b_1] \cup \{a_1, b_1\}, P_2 \cap [a_2, b_2] \cup \{a_2, b_2\}, ..., P_n \cap [a_n, b_n] \cup \{a_n, b_n\})$ is a grid on R where each element of H is contained in some element of G. Furthermore, $C \subseteq H$ since if $Q_t \in G$ and $Q_t \subseteq R$ then it must follow that $Q_t = \prod_{i=1}^{n} [x_{s_i-1}^{(j)}, x_{s_i}^{(j)}]$ for some choices of $1 \leq s_i \leq n_i$. Since $Q_t \subseteq R$ it follows that $[x_{s_i-1}^{(j)}, x_{s_i}^{(j)}] \subset [a_i, b_i]$ for each $i \in \{1, 2, 3, ..., n\}$, which means that $x_{s_i-1}^{(j)}, x_{s_i}^{(j)} \in P_i \cap [a_i, b_i]$ and therefore $Q_t \in H$. Hence, for each $Q_t \in C$ we know that $Q_t \in H \cap G$. It follows that $\sum_{Q_t \in C} |Q_t| \leq \sum_{Q_t \in H \cap G} |Q_t| \leq \sum_{K_i \in H} |K_i| = |R|$ by Theorem 12.19.

□

We now show that the volume $|R|$ if a rectangle is the same as the volume $Vol(R)$ of the rectangle.

Theorem 12.21. *Let $\mathcal{D} = \{R_1, R_2, ..., R_k\}$ be a collection of rectangles so that $\bigcup_{i=1}^{k} R_i^\circ$ covers rectangle R. Then $\sum_{i=1}^{k} |R_i| \geq |R|$, and $Vol(R) = |R|$.*

Proof. Since R is compact, by the Lebesgue Number Lemma we can find $\delta > 0$ so that if S is a set intersecting R of diameter less than δ then $S \subseteq R_i^\circ$ for some $i \in \{1, 2, 3, ..., k\}$. Let $G = G(P)$ be a grid on R where $P = \{P_1, P_2, ..., P_n\}$ and $P_i = \{x_0^{(i)}, x_1^{(i)}, ..., x_{n_i}^{(i)}\}$ for each $1 \leq i \leq n$, so that $|G| < \delta$. For each $i \in \{1, 2, 3, ..., k\}$, define $C_i = \{Q_t \in G | Q_t \subseteq R_i\}$. Then $\sum_{i=1}^{n} \sum_{Q_t \in C_i} |Q_t| \geq |R|$ since every $Q_t \in C_i$ for some i for every $Q_t \in G$, and $\sum_{Q_t \in G} |Q_t| = |R|$ by Theorem 12.19.

For each j, we know $\sum_{Q_t \in C_j} |Q_t| \leq |R_j|$ by Theorem 12.20, so it follows that $|R| \leq \sum_{i=1}^{n} \sum_{Q_t \in C_i} |Q_t| \leq \sum_{i=1}^{k} |R_i|$. Thus, $|R| \leq Vol(R)$ by Theorem 12.9.

Since $\{R\}$ is a cover of R, we know that $Vol(R) \leq |R|$. Hence, $Vol(R) = |R|$.

□

We next show that if you take any rectangle R containing a Jordan region E then the infimum of all of the outer sums of E with respect to grids on R is the same as the volume of the Jordan region. This lets us characterize Jordan region with grids, which is sometimes helpful. We observe that this is true regardless of the choice of rectangle R used, so it makes no difference to the infimum of the outer sums of E which rectangle containing E is used.

Theorem 12.22. *Let E be a Jordan region contained in a rectangle R in $\mathbb{R}^n$. Then $Vol(E) = \inf_G V(E,G)$, where G ranges over all grids on R.*

Proof. By definition, for any grid G on R, the outer sum $V(E,G) = \sum_{\{R_t \in G | R_t \cap \overline{E} \neq \emptyset\}} |R_t|$, which is a sum of volumes of the elements of a finite cover of E by rectangles, so $V(E,G) \geq Vol(E)$ for all grids G, which means that $\inf_G V(E,G) \geq Vol(E)$.

Let $\epsilon > 0$. We know that $\overline{E}$ is a Jordan region with $Vol(\overline{E}) = Vol(E)$ by Theorem 12.14. Let $\{R_i^\circ\}_{1 \leq i \leq k}$ be a finite cover of $\overline{E}$ by interiors of rectangles R_i so that $\sum_{i=1}^{k} |R_k| < Vol(E) + \epsilon$. By the Lebesgue Number Lemma we can find $\delta > 0$ so that if S is a set with $diam(S) < \delta$ and S intersects $\overline{E}$ then $S \subseteq R_i^\circ$ for some i. Choose a grid $H = \{Q_t\}_{1 \leq t \leq w}$ on R with $|H| < \delta$. For each $i \in \{1, 2, 3, ..., k\}$ let $C_i = \{Q_t \in H | Q_t \subseteq R_i\}$. By Theorem 12.20, we know that $\sum_{Q_t \in C_i} |Q_t| \leq |R_i|$, so $\sum_{i=1}^{k} \sum_{Q_t \in C_i} |Q_t| < Vol(E) + \epsilon$. Furthermore, if $Q_t \cap \overline{E} \neq \emptyset$ then $Q_t \subseteq R_i^\circ$ for some i, which means that $V(E,H) = \sum_{\{Q_t \in H | Q_t \cap \overline{E} \neq \emptyset\}} |Q_t| \leq \sum_{i=1}^{k} \sum_{Q_t \in C_i} |Q_t| < Vol(E) + \epsilon$. Hence, $\inf_G V(E,G) \leq Vol(E)$ and therefore $\inf_G V(E,G) = Vol(E)$.

□

Next, we show that, much as integrals exist when upper sum can be made arbitrarily close to lower sums, a region is a Jordan region (and volume exists) if and only if we can find grids where the inner and outer sums over those grids can be made as close as we wish.

Theorem 12.23. *Let $E \subseteq \mathbb{R}^n$. Then E is a Jordan region if and only if, for every $\epsilon > 0$ there is a grid G on a rectangle R containing E so that $V(E,G) - v(E,G) = V(\partial(E), G) < \epsilon$, in which case $Vol(E) = \sup_G v(E,G) = (I)Vol(E) = (O)Vol(E) = \inf_G V(E,G)$.*

Proof. We know that E is a Jordan region if and only if $\partial(E) = 0$ which is true if and only if for every $\epsilon > 0$ there is a grid G on a rectangle R containing E so that $V(E, \partial(E)) = \sum_{\{R_j \in G | R_j \cap \partial(E) \neq 0\}} |R_j| < \epsilon$. We know that $R_j \cap \partial(E) \neq 0$ if and only if $R_j \not\subseteq R^\circ$ by Theorem 12.15. Hence, $S(E,G) = O(E,G) \backslash I(E,G)$. Therefore, $V(E,G) - v(E,G) = \sum_{R_j \in O(E,G)} |R_j| - \sum_{R_j \in I(E,G)} |R_j| = \sum_{R_j \in S(\partial(E),G)} |R_j| = V(\partial(E), G)$. Thus, E is a Jordan region if and only if for every $\epsilon > 0$ we can find a grid G so that $V(\partial(E), G) < \epsilon$, which is true if and only if $V(E,G) - v(E,G) < \epsilon$.

Let E be a Jordan region and note that $Vol(E) = \inf_G V(E,G)$. By Theorem 12.16 we know that $v(E,G) \leq Vol(E)$ for every grid G. Let $\epsilon > 0$. Choose a grid G so that $V(E,G) < Vol(E) + \epsilon$ and a grid H so that $V(E,H) - v(E,H) < \epsilon$. Let K be a refinement of G and H. Then $V(E,K) - \epsilon < v(E,H) \leq Vol(E) \leq V(E,K) < Vol(E) + \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $Vol(E) = \sup_G v(E,G)$. □

The intersection, set difference and union of Jordan regions is a Jordan region, and the volume of the union of non-overlapping Jordan regions is the sum of the volumes. Combining this with earlier theorems helps us to start to use rectangle volumes and grids in a way that fits with our intuition. In particular, the sum of the volumes of rectangles in a grid is the volume of the union of the those rectangles. This gets us to the stage where we have rigorously established that volume works largely in the way that we would like it to since these grid rectangles can then be used to approximate the volumes of every other Jordan region.

Theorem 12.24. *Let E_1 and E_2 be Jordan regions. Then $E_1 \cup E_2$, $E_1 \cap E_2$ and $E_1 \setminus E_2$ are Jordan regions so that $Vol(E_1 \cup E_2) \leq Vol(E_1) + Vol(E_2)$, $Vol(E_1 \cap E_2) \leq Vol(E_1)$ and $Vol(E_1 \setminus E_2) \leq Vol(E_1)$. If E_1 and E_2 are non-overlapping then $Vol(E_1 \cup E_2) = Vol(E_1) + Vol(E_2)$.*

Proof. Let $\epsilon > 0$. We can find $\mathcal{C}_1 = \{R_1, R_2, ..., R_k\}$ and $\mathcal{C}_2 = \{T_1, T_2, ..., T_j\}$, collections of rectangles covering E_1 and E_2 respectively, so that $\sum_{i=1}^{k} |R_k| < Vol(E_1) + \frac{\epsilon}{2}$ and $\sum_{i=1}^{j} |T_j| < Vol(E_2) + \frac{\epsilon}{2}$. Then $\mathcal{C}_1 \cup \mathcal{C}_2$ is a finite collection of rectangles covering $E_1 \cup E_2$, the sum of whose volumes less than $Vol(E_1) + Vol(E_2) + \epsilon$. In particular, since $Vol(\partial(E_1) = 0$ and $Vol(\partial(E_2) = 0$ we can choose $\mathcal{C}_1 \cup \mathcal{C}_2$ to be a collection of rectangles, the sum of whose volumes is less than ϵ which means that $\partial(E_1) \cup \partial(E_2)$ is a Jordan region with $Vol(\partial(E_1) \cup \partial(E_2)) = 0$ by Theorem 12.14.

Since $\partial(E_1 \cup E_2) \subset \partial(E_1) \cup \partial(E_2)$ and $\partial(E_1 \cap E_2) \subset \partial(E_1) \cup \partial(E_2)$ and $\partial(E_1 \setminus E_2) \subseteq \partial(E_1) \cup \partial(E_2)$, it follows from Theorem 12.11 that $Vol(\partial(E_1 \cup E_2)) = 0$, $Vol(\partial(E_1 \cap E_2)) = 0$, and $Vol(\partial(E_1 \setminus E_2)) = 0$, so $E_1 \cup E_2$, $E_1 \cap E_2$ and $E_1 \setminus E_2$ are Jordan regions.

Since $E_1 \cap E_2 \subseteq E_1$ and $E_1 \setminus E_2 \subseteq E_1$ we know that $Vol(E_1 \cap E_2) \leq Vol(E_1)$ and $Vol(E_1 \setminus E_2) \leq Vol(E_1)$ by Theorem 12.11. Since $Vol(E_1 \cup E_2) \leq Vol(E_1) + Vol(E_2) + \epsilon$ for all $\epsilon > 0$ it follows that $Vol(E_1 \cup E_2) \leq Vol(E_1) + Vol(E_2)$.

Next, assume that E_1 and E_2 are non-overlapping. Then we know that $Vol(E_1 \cap E_2) = 0$. Note that if $\in E_1^\circ \cap E_2^\circ \neq \emptyset$ then this intersection contains a rectangle L, which means that $|L| < Vol(E_1 \cap E_2)$ by Theorem 12.16, which is impossible. Hence, $E_1 \cap E_2 \subset \partial(E_1) \cup \partial(E_2)$, which has volume zero.

Let $\epsilon > 0$. By Theorem 12.22, we can find a grid G_1 on R so that $V(\partial(E_1) \cup \partial(E_2), G_1) < \epsilon$. We can also find a grid G_2 on R so that $V(E_1 \cup E_2, G_2) < Vol(E_1 \cup E_2) + \epsilon$. Let H be any refinement of $G_1 * G_2$, a common refinement of G_1 and G_2. Then by Theorem 12.3, it follows that $V(E_1, E_2, H) < Vol(E_1 \cup E_2) + \epsilon$ and also $V(\partial(E_1) \cup \partial(E_2)) < \epsilon$.

Let $C_1 = \{W_t \in H | W_t \subseteq E_1^\circ\}$, $C_2 = \{W_t \in H | W_t \subseteq E_2^\circ\}$ and $C_3 = \{W_t \in H | W_t \cap (\partial(E_1) \cup \partial(E_2)) \neq \emptyset\}$. By Theorem 12.15, we know that each rectangle in H which intersects E_1 either intersects the boundary of E_1 is is contained in the interior of E_1. Hence, $C_1 \cup C_3$ covers E_1, which means that $\sum_{W_t \in C_1} |W_t| + \sum_{W_t \in C_3} |W_t| \geq Vol(E_1)$. Since $\sum_{W_t \in C_3} |W_t| < \epsilon$, it follows that $\sum_{W_t \in C_1} |W_t| \geq Vol(E_1) - \epsilon$. By a similar argument, $\sum_{W_t \in C_2} |W_t| \geq Vol(E_2) - \epsilon$. Since $\in E_1^\circ \cap E_2^\circ = \emptyset$, it follows that $C_1 \cap C_2 = \emptyset$ and therefore $V(E, H) = \sum_{\{W_t \in H | W_t \cap (\overline{E_1 \cup E_2}) \neq \emptyset\}} |W_t| \geq$

$$\sum_{W_t\in C_1}|W_t|+\sum_{W_t\in C_2}|W_t|\geq Vol(E_1)+Vol(E_2)-2\epsilon.$$

The partition H could have been chosen to refine any grid, and ϵ could have been any positive number, which means that if G is a grid on R then for each $\epsilon>0$ we can choose a grid H refining G so that $V(E,H)\leq V(E,G)$ and $V(E,H)\geq Vol(E_1)+Vol(E_2)-2\epsilon$ and therefore $V(E,G)\geq Vol(E_1)+Vol(E_2)-2\epsilon$ for every $\epsilon>0$, and thus $V(E,G)\geq Vol(E_1)+Vol(E_2)$. Since this is true for all grids G on R we know that $Vol(E_1\cup E_2)=\inf_G V(E,G)\geq Vol(E_1)+Vol(E_2)$. It follows that $Vol(E_1\cup E_2)=Vol(E_1)+Vol(E_2)$.

□

Definition 91

Let $f:E\to\mathbb{R}$ be bounded, E a Jordan region. We define the *characteristic function* of a set S so be $\chi_S(\mathbf{x})=1$ if $\mathbf{x}\in S$ and $\chi_S(\mathbf{x})=0$ otherwise. We define the *zero extension* of f to R to be the function $F(x)=\chi_E(\mathbf{x})f(\mathbf{x})$. We say that f is integrable on E if the zero extension of F of f is integrable on R, and define $\int_E f=\int_R F$. If $E\subset V$, another Jordan region (but f is only defined on E) then if R is a rectangle containing V we also define $\int_V f=\int_E f=\int_R F$. We also define the *zero boundary* extension of f to E to be $G(\mathbf{x})=\chi_{E^\circ}(\mathbf{x})f(\mathbf{x})$, the function which is f on the interior of E, but zero on the boundary of E and outside of E.

We next show that any bounded function on a domain of volume zero is integrable, and has an integral of zero.

Theorem 12.25. *Let $E\subset\mathbb{R}^n$ so that $Vol(E)=0$, and let $f:E\to\mathbb{R}^n$ be a bounded function. Then $\int_E f=0$.*

Proof. Let $M>|f(\mathbf{x})|$ for all $\mathbf{x}\in E$. Since $Vol(E)=0$ there is a grid $G=\{R_i\}_{1\ leqi\leq k}$ on a rectangle containing E so that $V(E,G)<\frac{\epsilon}{2M}$ by Theorem 12.22, which means that $|\sum_{R_j\in S(G)}M_j|R_j||\leq M\sum_{R_j\in O(G)}|M_j|<\frac{\epsilon}{2}$, and $|\sum_{R_j\in S(G)}m_j|R_j||\leq M\sum_{R_j\in O(G)}|m_j|<\frac{\epsilon}{2}$. Thus, $U(f,G)-L(f,G)<\epsilon$ so f is integrable, and $-\frac{\epsilon}{2}<L(f,G)\leq\int_E f\leq U(f,G)<\frac{\epsilon}{2}$. Since this is true for all $\epsilon>0$ we know that $\int_E f=0$. □

It can be helpful in some instances to know that if we can cover a finite set of rectangles with a countable collection the sum of whose volumes is small, then the volume of the union of the finite set of rectangles is also small.

Theorem 12.26. *Let $F = \{R_i\}_{1\leq i\leq k}$ be a finite collection of non-overlapping rectangles in $\mathbb{R}^n$, and let $\epsilon > 0$. Let $\{I_i^\circ\}_{i\in\mathbb{N}}$ be a cover of $\bigcup F$, where each I_i is also a rectangle in $\mathbb{R}^n$ and $\sum_{i=1}^{\infty} |I_i| < \epsilon$. Then $\sum_{i=1}^{n} |R_i| < \epsilon$.*

Proof. Since $\bigcup F$ is a finite union of rectangles, each of which is compact, $\bigcup F$ is closed and bounded and therefore compact by the Heine-Borel Theorem. By the Lebesgue Number Lemma there is a $\delta > 0$ so that if S is a set with $diam(S) < \delta$ and $S \cap (\bigcup F) \neq \emptyset$ then $S \subseteq I_i^\circ$ for some natural number I. Choose grids G_i on R_i for $1 \leq i \leq k$ so that $|G_i| < \delta$ for each $i \in \{1, 2, 3, ..., k\}$, and let $W = \bigcup_{i=1}^{k} G_i$. For each natural number i, let $C_i = \{Q_t \in W | Q_t \subseteq I_i\}$. Observe that C_i is a collection of non-overlapping rectangles because if $Q_t \in G_m$ and $Q_s \in G_r$ then $Q_t \cap Q_s \subseteq R_m \cap R_r$, which has volume zero since R_m and R_s are non-overlapping, and if Q_t, Q_s are contained in the same grid G_m then Q_t, Q_s are non-overlapping because elements of a grid on a rectangle do not overlap.

Since there are only finitely many elements in W, there is a last integer j so that $C_j \neq \emptyset$. Since $|Q_t| < \delta$ for all $Q_t \in W$ we know that each Q_t is contained in some I_i, which means that $W = \bigcup_{i=1}^{j} C_i$. We know that $Vol(\bigcup C_i) \leq I_i$ for each $i \in \{1, 2, 3, ..., j\}$ since $\bigcup C_i \subseteq I_i$. Since the elements of C_i are non-overlapping, we know that $Vol(\bigcup C_i) = \sum_{Q_t\in C_i} |Q_t| = \sum_{\{Q_t\in W|Q_t\subseteq I_i\}} |Q_t| < |I_i|$. From this, it follows that $\sum_{Q_t\in W} |Q_t| = \sum_{i=1}^{k} |R_i| < \sum_{i=1}^{j} |I_i| < \sum_{i=1}^{\infty} |I_i| < \epsilon$.

□

Next we show that set has volume zero if and only we can cover the set by finite collections of cubes the sum of whose volumes can be made arbitrarily small. The advantage to being able to use cubes is that under certain nice mappings we can make the images of cubes behave better than arbitrary rectangles (the volume of the domain and range are more easily comparable, for instance).

Theorem 12.27. *Let $E \subset \mathbb{R}^n$. Then $Vol(E) = 0$ if and only if for every $\epsilon > 0$ there is a finite collection $F = \{C_i\}_{1\leq i\leq k}$ of cubes of equal side length whose interiors cover $\overline{E}$, the sum of whose volumes is less than ϵ. Furthermore, for every $\zeta > 0$ we can choose such a collection F so that the side length of the elements of F is less than ζ.*

Proof. If such a collection of cubes exists for every $\epsilon > 0$ then by definition $Vol(E) = 0$.

Assume $Vol(E) = 0$. Then by Theorems 12.14 and 12.9 we can find rectangles $\{R_i\}_{1\leq i\leq t}$ so that $\overline{E} \subset \bigcup_{i=1}^{t} R_i^\circ$ and $\sum_{i=1}^{k} |R_i| < \frac{\epsilon}{2}$.

Let $\zeta > 0$. By the Lebesgue Number Lemma we can find a $\delta > 0$ so that if a set S has diameter less than δ and $S \cap \overline{E} \neq \emptyset$ then $S \subset R_i^\circ$ for some $i \in \{1, 2, 3, ..., t\}$. Choose a cube

Q containing $\bigcup_{i=1}^{t} R_i$, and a partition P for Q consisting of equally spaced partition points in each edge factor of Q inducing a grid G whose elements are cubes $\{Q_i\}_{1\leq i\leq s}$ of equal side length $L < \zeta$ so that $|G| = \sqrt{n}L < \frac{\delta}{2}$. Let $T = \{Q_i \in G | Q_i \cap \overline{E} \neq \emptyset\} = \{Q_{n_1}, Q_{n_2}, ..., Q_{n_k}\}$. By Theorem 12.26, we know that $\sum_{i=1}^{k} |Q_{n_i}| < \frac{\epsilon}{2}$.

By Theorem 12.7 we can find $\gamma \in (L, \zeta)$ so that for each $Q_{n_i} \in T$, we can find a cube W_i of side length γ so that $Q_{n_i} \subset W_i^\circ$ and $|W_i| - |Q_{n_i}| < \frac{\epsilon}{2k}$. Then $\overline{E} \subseteq \bigcup_{i=1}^{k} Q_{n_i} \subset \bigcup_{i=1}^{k} W_i^\circ$ and $\sum_{i=1}^{k} |W_i| = \sum_{i=1}^{k} |Q_{n_i}| + \sum_{i=1}^{k} |W_i| - |Q_{n_i}| < \frac{\epsilon}{2} + k\frac{\epsilon}{2k} = \epsilon$.

□

Next, we show that Riemann sums can be made as close as we wish to upper and lower sums.

Theorem 12.28. *Let $f : R \to \mathbb{R}$ be bounded, where R is a rectangle, and let $\epsilon > 0$ and let G be a grid on R. Then there are markings T, R of G so that $U(f, G) - S_T(f, G) < \epsilon$ and $S_R(f, G) - L(f, G) < \epsilon$.*

Proof. Since $M_t = \sup_{x\in R_t} f(\mathbf{x})$ and $m_t = \inf_{x\in R_t} f(\mathbf{x})$, we can find points $\mathbf{r}_t^*, \mathbf{s}_t^* \in R_t$ so that $M_t - f(\mathbf{s}_t^*) < \frac{\epsilon}{|R|}$ and $f(\mathbf{r}_t^*) - m_t < \frac{\epsilon}{|R|}$. This gives us markings $T = \{\mathbf{s}_t^* | R_t \in G\}$, $R = \{\mathbf{r}_t^* | R_t \in G\}$ so that $U(f, G) - S_T(f, G) < \epsilon$ and $S_R(f, G) - L(f, G) < \epsilon$.

□

We parallel theorems in single variable integration by next showing that if a function is integrable then a grid with sufficiently small mesh will have upper and lower sums that are as close together as we wish. This is an "if and only if" condition but the other direction has already been proven (we already know that if the upper and lower sums can be made as close as we wish then the function is integrable).

Theorem 12.29. *Let R be a rectangle in $\mathbb{R}^n$ and let E be a Jordan region contained in R. Let $f : E \to \mathbb{R}$ be integrable. Then for any $\epsilon > 0$ there is a $\delta > 0$ so that if G is a grid on R so that $|G| < \delta$ then $U(f, G) - L(f, G) < \epsilon$.*

Proof. Since $\epsilon > 0$ we can find a grid G on R so that $U(f, G) - L(f, G) < \frac{\epsilon}{2}$ and $G = \{R_1, R_2, ..., R_k\}$. Since f is bounded we can find $M > 0$ so that $|f(\mathbf{x})| < M$ for all $\mathbf{x} \in E$. Let $S = \bigcup_{R_i\in G} \partial(R_i)$. Then the the volume of S is zero, so we can find a finite covering of S by cubes $\{Q_1, Q_2, ..., Q_m\}$ of equal side length s so that $\sum_{i=1}^{m} |Q_i| < \frac{\epsilon}{4M}$. By increasing

s slightly we obtain cubes $\{B_1, B_2, ..., B_m\}$ of side length $t > s$ so that $Q_i \subset B_i^\circ$ for each i and $\sum_{i=1}^{m} |B_i| < \frac{\epsilon}{4M}$. Thus, the $\{B_1^\circ, B_2^\circ, ..., B_m^\circ\}$ are an open covering S. By the Lebesgue Number Lemma we can find a $\delta > 0$ so that if T is a set of diameter less than δ then if $T \cap S \neq \emptyset$ it follows that $T \subset B_i^\circ$ for some i.

Let H be a grid on R with mesh $|H| < \delta$, where $H = \{T_1, T_2, ..., T_w\}$. For each $T_i \in H$ and $R_i \in G$ we know that exactly one of $T_j \cap R_j = \emptyset$, $T_j \subseteq R_j^\circ$ or $T_j \cap \partial(R_j) \neq \emptyset$ is true by Theorem 12.15. Thus, $U(f, H) - L(f, H) = \sum_{\{T_j \in H | T_j \cap S = \emptyset\}} (M_j^H - m_j^H)(|T_j|)$ $+ \sum_{\{T_j \in H | T_j \cap S \neq \emptyset\}} (M_j^H - m_j^H)(|T_j|)$.

For the first sum, we note that for each T_j where $T_j \cap S = \emptyset$, we know that $T_j \subset R_i^\circ$ for some $R_i \in G$, and so $M_j^H - m_j^H \leq M_i^G - m_i^G$. Since $\sum_{\{T_j \in H | T_j \subset R_i}} |T_j| \leq |R_i|$ it follows that

$$\sum_{\{T_j \in H | T_j \cap S = \emptyset\}} (M_j^H - m_j^H)(|T_j|) \leq \sum_{i=1}^{k} (M_i^G - m_i^G)|R_i| = U(f, G) - L(f, G) < \frac{\epsilon}{2}.$$

For the second sum, we know that each $\sum_{\{T_j \in H | T_j \cap S \neq \emptyset\}} |T_j| = Vol(\bigcup_{\{T_j \in H | T_j \cap S \neq \emptyset\}} T_j)$ since the T_j rectangles are non-overlapping. Since each T_j rectangle intersecting S is a subset of some B_i we know that $Vol(\bigcup_{\{T_j \in H | T_j \cap S \neq \emptyset\}} T_j) \leq Vol(\bigcup_{i=1}^{m} B_i) \leq \sum_{i=1}^{m} |B_i| < \frac{\epsilon}{4M}$. From this, we conclude that $\sum_{\{T_j \in H | T_j \cap S \neq \emptyset\}} (M_j^H - m_j^H)(|T_j|) \leq 2M(\frac{\epsilon}{4M}) = \frac{\epsilon}{2}$. Hence $U(f, H) - L(f, H) < \epsilon$.

□

We show that if we take any sequence of grids whose meshes go to zero (grids induced by standard partitions on the edge factors of a cube, for instance) then a function integrable if and only if, regardless of marking, the Riemann sums always converge to the same value (in which case that value is the integral).

Theorem 12.30. *Let E be a Jordan region contained in rectangle $R \subset \mathbb{R}^n$. Let $f : E \to \mathbb{R}$ be bounded. Let $\{G_n\}$ be a sequence of grids on R so that $\{|G_n|\} \to 0$. Let F be the zero extension of f to R. Then f is integrable if and only if there is a number I so that $\{S_{T_n}(F, G_n)\} \to I$ regardless of choice of markings T_n of G_n, in which case $\int_E f = I$.*

Proof. First, assume that f is integrable and $\int_E f = I$. Let $\epsilon > 0$. By Theorem 12.29, we can find $\delta > 0$ so that if G is a grid on R and $|G| < \delta$ then $U(F, G) - L(F, G) < \epsilon$.

Choose $k \in \mathbb{N}$ so that if $n \geq k$ then $|G_n| < \delta$. Then if $n \geq k$ we know that $U(F, G_n) - L(f, G_n) < \epsilon$. Since, for any marking T_n it is true that $S_{T_n}(F, G_n), I \in [L(f, G_n), U(f, G_n)]$ it follows that $|S_{T_n}(F, G_n) - I| < \epsilon$. Thus, $S_{T_n}(F, G_n) \to I$ regardless of choice of markings T_n of G_n.

Next, assume that $S_{T_n}(F,G_n) \to I$ regardless of choice of markings T_n of G_n. For each $n \in \mathbb{N}$, by Theorem 12.28, we can choose markings T_n, R_n of G_n so that $U(F,G_n) - S_{T_n}(F,G_n) < \frac{1}{n}$ and $S_{R_n}(F,G_n) - L(F,G_n) < \frac{1}{n}$. Since $\{S_{T_n}(F,G_n)\} \to I$ and $\{U(F,G_n) - S_{T_n}(F,G_n)\} \to 0$ we know $\{U(F,G_n)\} \to I$. Since $\{S_{R_n}(F,G_n)\} \to I$ and $\{S_{T_n}(F,G_n) - L(F,G_n)\} \to 0$ we know $\{L(F,G_n)\} \to I$ and $\{U(F,G_n) - L(F,G_n)\} \to 0$. Hence, given any $\epsilon > 0$ we can choose k so that if $n \geq k$ then $U(F,G_n) - L(F,G_n) < \epsilon$ and $|S_{T_n}(F,G_n) - I| < \epsilon$, so F is integrable on R which means f is integrable on E. Since $\int_R F, S_{T_n}(F,G_n) \in [L(F,G_n), U(F,G_n)]$ if follows that $|\int_R F - S_{T_n}(F,G_n)| < \epsilon$ and since $|S_{T_n}(F,G_n) - I| < \epsilon$, we know that $|I - \int_R F| < 2\epsilon$. Since this is true for all $\epsilon > 0$, it follows that $I = \int_R F = \int_E f$. □

We next develop a way to determine whether a function is integrable on a Jordan region based on the measure of the discontinuities of the function. This takes a few steps.

Definition 92

Let $f : E \to \mathbb{R}$ be bounded, E a Jordan region. For each set S in $\mathbb{R}^n$ which intersects E we define the *oscillation of f on S* to be $\Omega_f(S) = \sup_{\mathbf{x} \in S \cap E} f(\mathbf{x}) - \inf_{\mathbf{x} \in S \cap E} f(\mathbf{x})$. If $\mathbf{p} \in \overline{E}$ then we define the *oscillation of f at $\boldsymbol{p}$* to be $\omega_f(\mathbf{p}) = \inf_{\epsilon > 0} \Omega_f B_\epsilon(\mathbf{p}) = \lim_{h \to 0^+} \Omega_f B_\epsilon(\mathbf{p}) = \inf_{R_\epsilon} \Omega_f(R_\epsilon)$, where R_ϵ is an n-rectangle containing $\mathbf{p}$ in its interior whose diameter is ϵ, and $\epsilon > 0$.

We could have defined integrability on arbitrary regions rather than just Jordan regions in the manner described in the definition of integrability over a Jordan region, but for continuous functions that are non-zero on their boundaries the set of regions for which the function would have been integrable would just be exactly the Jordan regions anyway, and Jordan regions behave nicely under maps like those described in the Inverse Function Theorem (maps that are one to one, continuously differentiable and have a derivative with non-zero determinant).

It sometimes looks neater to write $\Omega_f S$ rather than $\Omega_f(S)$, but both mean the same thing. For instance, on an open interval we will write $\Omega_f(a,b)$ instead of $\Omega_f((a,b))$. We should also prove that $\inf_{\epsilon > 0} \Omega_f B_\epsilon(p) = \lim_{h \to 0^+} \Omega_f B_\epsilon(p) = \inf_{R_\epsilon} \Omega_f(R_\epsilon)$ rather than simply claim that these are all the same, so we do this below.

Theorem 12.31. *$f : D \to \mathbb{R}$ be bounded and let I_1, I_2 be sets intersecting D with $I_1 \subseteq I_2$. Then $\Omega_f(I_1) \leq \Omega_f(I_2)$.*

Proof. We know $\sup_{x \in I_1 \cap D} f(x) \leq \sup_{x \in I_2 \cap D} f(x)$ and $\inf_{x \in I_1 \cap D} f(x) \geq \inf_{x \in I_2 \cap D} f(x)$, which means $\Omega_f(I_1) \leq \Omega_f(I_2)$. □

Theorem 12.32. *Let $f : D \to \mathbb{R}$ be bounded and let $p \in \overline{D}$. Then there is a non-negative number $\omega_f(p) = \inf\limits_{\epsilon>0} \Omega_f(B_\epsilon(p)) = \lim\limits_{h\to 0^+} \Omega_f(B_h(p)) = \inf\limits_{R_\epsilon} \Omega_f(R_\epsilon)$ where R_ϵ is any rectangle of diameter ϵ containing p in its interior and $\epsilon > 0$.*

Proof. Note that $\Omega_f(S)$ is a non-negative real number for each set S which intersects D, so $w = \inf\limits_{\gamma>0} \Omega_f(B_\gamma(p))$ exists and is non-negative. Let $\epsilon > 0$. Then for some $\delta > 0$ we know that $\Omega_f(B_\delta(p)) < w + \epsilon$ by the approximation property. However, we also know that if $0 < h < \delta$ then $\Omega_f(B_h(p)) \leq \Omega_f(B_\delta(p))$ by Theorem 12.31, so $w = \omega_f(p) = \lim\limits_{h\to 0^+} \Omega_f(B_h(p))$. Similarly, if R has diameter less than δ and contains p in its interior then $\Omega_f(R) < w + \epsilon$ since $R \subset B_\delta(p)$. Likewise, if R is any rectangle containing p in its interior then there is a δ' so that $B_{\delta'}(p) \subset R$, so $\Omega_f(B_{\delta'}(p)) \leq \Omega_f(R)$. Hence, $\inf\limits_{R_\epsilon} \Omega_f(R_\epsilon) = \inf\limits_{\epsilon>0} \Omega_f B_\epsilon(p) = w_f(p)$. □

Theorem 12.33. *Let $f : E \to \mathbb{R}$ be bounded and let $\epsilon > 0$, where $E \subseteq \mathbb{R}^n$, and let $\boldsymbol{p} \in E^\circ$ where $\omega_f(\boldsymbol{p}) \geq \epsilon$. Then $\Omega_f E \geq \epsilon$.*

Proof. Choose $\delta > 0$ so that $B_\delta(\mathbf{p}) \subseteq E$. Thus, by theorem 12.31 and the definition of oscillation at a point, $\Omega_f(E) \geq \Omega_f(B_\delta(\mathbf{p})) \geq \omega_f(\mathbf{p}) \geq \epsilon$.

□

Continuity can be characterized in terms of oscillation. We are working towards proving that a function on a Jordan region is integrable if and only if its set of discontinuities has Lebesge measure zero. In the next theorem we show that a function is continuous at a point if and only if its oscillation at that point is zero.

Theorem 12.34. *Let $f : D \to \mathbb{R}$ be bounded and let $p \in D$. Then f is continuous at p if and only if $\omega_f(p) = 0$.*

Proof. Assume f is continuous at p and let $\epsilon > 0$. Choose $\delta > 0$ so that if $|\mathbf{x} - \mathbf{p}| < \delta$ and $\mathbf{x} \in D$ then $|f(\mathbf{x}) - f(\mathbf{p})| < \frac{\epsilon}{2}$. Then $\sup\limits_{x\in B_\delta(\mathbf{p})\cap D} f(\mathbf{x}) \leq f(\mathbf{x}) + \frac{\epsilon}{2}$ and $\inf\limits_{x\in B_\delta(\mathbf{p})\cap D} f(\mathbf{x}) \geq f(\mathbf{x}) - \frac{\epsilon}{2}$. Thus $\Omega_f B_\delta(\mathbf{p}) \leq \epsilon$ so $\omega_f(\mathbf{p}) \leq \epsilon$ for all $\epsilon > 0$, which means that $\omega_f(\mathbf{p}) = 0$.

Assume that $\omega_f(\mathbf{p}) = 0$ and let $\epsilon > 0$. Then since $\omega_f(\mathbf{p}) = \inf\limits_{h\in\mathbb{R}^+} \Omega_f B_h(\mathbf{p})$, by the Approximation Property we can find $\delta > 0$ so that $\Omega_f(p - \delta, p + \delta) < \epsilon$, which means that if $|\mathbf{x} - \mathbf{c}| < \delta$ and $x \in D$ then $|f(\mathbf{x}) - f(\mathbf{c})| < \epsilon$, so f is continuous at c. □

We have more control over compact set behavior most of the time. While the set of discontinuities may not be compact for a function, we can show that the set of points in a compact domain where the oscillation is more than or equal to any particular value is compact, which is out objective below. Then we can describe the set of discontinuities of a function as a union of such sets.

Theorem 12.35. *Let K be a closed set in $\mathbb{R}^n$, and $f : K \to \mathbb{R}$ be bounded and $\epsilon > 0$, and let $E = \{\boldsymbol{x} \in K | \omega_f(\boldsymbol{x}) \geq \epsilon\}$. Then E is closed. If K is compact then E is compact.*

Proof. Let $\{\mathbf{x}_n\} \subseteq E$, where $\{\mathbf{x}_n\} \to \mathbf{x}$. Then for any $h > 0$ we know that $B_h(\mathbf{p})$ contains $\mathbf{x}_m$ for some $m \in \mathbb{N}$. Choose $\delta > 0$ we know that $B_\delta(\mathbf{x}_m) \subset B_h(\mathbf{p})$. Since $\epsilon \leq \omega_f(\mathbf{x}_m) \leq \Omega_f B_h(\mathbf{p})$ by Theorems 12.32 and 12.31, it follows that $\omega_f(\mathbf{p}) \geq \epsilon$. Hence, E contains all of its limit points and is closed. If K is compact then E is also bounded and thus compact by the Heine-Borel Theorem. □

We next show that if the oscillation at a point is small then there must be radii so that open balls with radii that small and therefore diameters so that rectangles with diameters that small containing the point on which the oscillation is small. This is addressed below.

Theorem 12.36. *Let $f : D \to \mathbb{R}$ be bounded and let $p \in \overline{D}$ and $\epsilon > 0$. If $\omega_f(\boldsymbol{p}) < \epsilon$ them there is a $\delta > 0$ so that $\Omega_f B_\delta(\boldsymbol{p}) < \epsilon$.*

Proof. This follows directly from the Approximation Property since we know that $\omega_f(\mathbf{p}) = \inf_{\{h \in \mathbb{R} | h > 0\}} \Omega_f B_h(\mathbf{p})$. □

The following theorem tells us that in a compact set if the oscillation is small at points of the set then the oscillation on small enough rectangles intersecting the set can also be made small.

Theorem 12.37. *Let $f : D \to \mathbb{R}$ be a bounded function, with K a compact subset of D. Let $\epsilon > 0$ and $\omega_f(\boldsymbol{p}) < \epsilon$ for each $\mathbf{p} \in K$. Then there is a $\gamma > 0$ so that if I is a set with diameter no more than γ and $I \cap K \neq \emptyset$ then $\Omega_f I < \epsilon$.*

Proof. By theorem 12.36 for each $\mathbf{p} \in K$ we can find an $\epsilon_\mathbf{p} > 0$ so that $\Omega_f(B_{\epsilon_\mathbf{p}}(\mathbf{p})) < \epsilon$. Then $\mathcal{C} = \{B_{\epsilon_\mathbf{p}}(\mathbf{p})\}_{p \in K}$ is an open cover of K, so by the Lebesgue Number Lemma we can find $\gamma > 0$ so that if I is a set so that $I \cap K \neq \emptyset$ and $\operatorname{diam}(I) \leq \gamma$ then $I \subseteq B_{\epsilon_\mathbf{p}}(\mathbf{p})$ for some $p \in K$, which means that $\Omega_f(I) < \epsilon$. □

Theorem 12.38. *Let E be a set in $\mathbb{R}^n$ so that for any countable collection of rectangles $\{R_i\}_{i \in \mathbb{N}}$ covering E, $\sum_{i=1}^{\infty} |R_i| \geq \gamma$. Then if $\{I_i\}_{i \in \mathbb{N}}$ is a collection of rectangles that covers E, it is also true that $\sum_{\{i | I_i^0 \cap E \neq \emptyset\}} |I_i| \geq \gamma$.*

Proof. Suppose that $\sum_{\{i | I_i^0 \cap E \neq \emptyset\}} |I_i| = \alpha < \gamma$ for a countable collection of rectangles $\{I_i\}_{i \in \mathbb{N}}$ covering E. Then all of the points of E not covered by $C = \{I_i | I_i^0 \cap E \neq \emptyset\}$ are contained in the boundaries of the I_i rectangles, each of which has Jordan content zero and therefore Lebesgue measure zero. Thus, if we set $B = \bigcup_{i=1}^{\infty} \partial(I_i)$ then $\lambda(B) = 0$ and so we can find a collection of rectangles $D = \{R_i\}_{i\in\mathbb{N}}$ which covers B so that $\sum_{i=1}^{\infty}|R_i| < \gamma - \alpha$. Hence, the set of all rectangles in $C \cup D$ is a countable collection of rectangles which cover E, the sum of whose volumes is less than γ, a contradiction.

□

The following theorem is the Lebesgue Characterization of Riemann Integrability in Euclidean spaces, though that is not the most common name for it. This condition frequently makes it easier to determine whether a function is integrable. A function on a Jordan region is integrable if and only if the set of points at which the function is not continuous is a set of Lebesgue measure zero.

Theorem 12.39. *Lebesgue Characterization of Riemann Integrability in $\mathbb{R}^n$. Let $f_E : E \to \mathbb{R}$ be bounded, where E is a Jordan region in $\mathbb{R}^n$ contained in the n-rectangle R, and let f be the zero extension of f_E to R. Then f_E is integrable on E (or equivalently, f is integrable on R) if and only if the set $D = \{\boldsymbol{x} \in R | f \text{ is not continuous at } \boldsymbol{x}\}$ has Lebesgue measure zero, which is true if and only if $W = \{\boldsymbol{x} \in E^\circ | f \text{ is not continuous at } \boldsymbol{x}\} = \{\boldsymbol{x} \in E^\circ | f_E \text{ is not continuous at } \boldsymbol{x}\}$ has Lebesgue measure zero.*

Proof. Since f is bounded we can choose $M > 0$ so that $|f(\mathbf{x})| < M$ for all $\mathbf{x} \in R$. For each $n \in \mathbb{N}$ let $D_n = \{\mathbf{x} \in R | \omega_f(\mathbf{x}) \geq \frac{1}{n}\}$. Note that $D = \bigcup_{n=1}^{\infty} D_n$. If $\lambda(D) = 0$ then $\lambda(D_n) = 0$ for each $n \in \mathbb{N}$ by Theorem 12.11.

Assume that $\lambda(D) \neq 0$. Then for some $m \in \mathbb{N}$ we know from theorem 12.13 that $\lambda(D_m) \neq 0$, so there is a number $\gamma > 0$ so that if $\{R_i\}_{i\in\mathbb{N}}$ is cover of E_m by rectangles then $\sum_{i=1}^{\infty}|R_i| \geq \gamma$. Let G be a grid on R. Then $U(f, G) - L(f, G) = \sum_{\{R_i \in G | R_i^\circ \cap D_m \neq \emptyset\}} (M_i - m_i)|R_i| + \sum_{\{R_i \in G | R_i^\circ \cap D_m = \emptyset\}} (M_i - m_i)|R_i|$. By Theorem 12.33 we know $M_i - m_i \geq \frac{1}{m}$ if $R_i^\circ \cap E_m \neq \emptyset$, so we know that $\sum_{\{R_i \in G | R_i^\circ \cap D_m \neq \emptyset\}} (M_i - m_i)|R_i| \geq \frac{\gamma}{m}$ by theorem 12.38, and thus f is not integrable.

Assume that $\lambda(D) = 0$. Let $\epsilon > 0$. Choose $j \in \mathbb{N}$ so that $\frac{|R|}{j} < \frac{\epsilon}{2}$. Choose a countable cover of D_j by interiors of rectangles $\mathcal{C} = \{I_i^\circ\}_{i\in\mathbb{N}}$ so that $\sum_{i=1}^{\infty}|I_i| < \frac{\epsilon}{4M}$. Let $K = R \setminus \bigcup \mathcal{C}$. We know K is compact by the Heine-Borel Theorem, and $\omega(p) < \frac{1}{j}$ for all $p \in K$, so by theorem 12.36 we can find a number $\delta > 0$ so that if I is a rectangle intersecting K with diameter less than δ then $\Omega_f(I) < \frac{1}{j}$.

Let $G = \{R_i\}_{1\leq i\leq s}$ be a grid on R with $|G| < \delta$. Then $U(f, G) - L(f, G) = \sum_{\{R_i \in G | R_i \cap K = \emptyset\}} (M_i -$

$m_i)|R_i| + \sum_{\{R_i \in G | R_i \cap K \neq \emptyset\}} (M_i - m_i)|R_i|$. Since the mesh of G is less than δ we have $\sum_{\{R_i \in G | R_i \cap K \neq \emptyset\}} (M_i - m_i)|R_i| < \frac{|R|}{j} < \frac{\epsilon}{2}$. Likewise, by Theorem 12.26, we know $\sum_{\{R_i \in G | R_i \cap K = \emptyset\}} |R_i| < \frac{\epsilon}{4M}$, it follows that $\sum_{\{R_i \in G | R_i \cap K = \emptyset\}} (M_i - m_i)|R_i| < 2M\frac{\epsilon}{4M} = \frac{\epsilon}{2}$. Hence, $U(f, G) - L(f, G) < \epsilon$ and f is integrable.

Finally, if $\mathbf{p} \in R \backslash \overline{E}$ then there is a $\delta > 0$ so $B_\delta(\mathbf{p}) \cap \overline{E} = \emptyset$ and therefore $f(\mathbf{x}) = f(\mathbf{p}) = 0$ if $\mathbf{x} \in B_\delta(\mathbf{p}) \cap R$, which means that f is continuous on $R \setminus \overline{E}$, so $D \subseteq (W \cup \partial(E))$. Since E is bounded we know that $\partial(E)$ is closed and bounded and therefore compact, which means that $Vol(E) = 0$ if and only if $\lambda(E) = 0$ by Theorem 12.10. Since E is a Jordan region, we know that $\lambda(\partial(E)) = Vol(\partial(E)) = 0$. Thus, if $\lambda(W) = 0$ then $\lambda(W \cup \partial(E)) = 0$, so $\lambda(D) = 0$ and f is integrable. If $\lambda(W) \neq 0$ then $\lambda(D) \neq 0$, so f is integrable.

□

The following theorem helps to motivate the reason for choosing Jordan regions as the regions over which we will consider integrals. Specifically, a constant function can only be integrated over a region which is a Jordan region.

Theorem 12.40. *Let $E \subseteq R$, a rectangle in $\mathbb{R}^n$. Then χ_E is integrable on R if and only if E is a Jordan region.*

Proof. Since E is bounded, $\partial(E)$ is closed and bounded and therefore compact by the Heine-Borel Theorem. Hence, $Vol(\partial(E)) = 0$ if and only if $\lambda(\partial(E)) = 0$ by Theorem 12.10. Let $\mathbf{p} \in E^\circ$. Then there is some $\delta > 0$ so that $B_\delta(\mathbf{p}) \subseteq E$, so $\chi_E(\mathbf{x}) = 1 = \chi_E(\mathbf{p})$ and thus $|\chi_E(\mathbf{x}) - \chi_E(\mathbf{p})| = 0$ if $|\mathbf{x} - \mathbf{p}| < \delta$. Hence, χ_E is continuous on E°.

Likewise, if $\mathbf{p} \in R \setminus \overline{E}$ then there is a $\delta > 0$ so $B_\delta(\mathbf{p}) \cap \overline{E} = \emptyset$ and therefore $\chi_E(\mathbf{x}) = \chi_E(\mathbf{p}) = 0$ if $\mathbf{x} \in B_\delta(\mathbf{p}) \cap R$, which means that χ_E is continuous on $R \setminus \overline{E}$.

However, if $\mathbf{p} \in \partial(E) \cap R^\circ$ then for every $\delta > 0$, $B_\delta(\mathbf{p})$ contains a point $\mathbf{x}_1 \in E$ and a point $\mathbf{x}_2 \in R \setminus E$, which means that $|\chi_E(\mathbf{x}_1) - \chi_E(\mathbf{x}_2)| = 1$ and therefore χ_E is not continuous at $\mathbf{p}$. Thus, if $E \subset R^\circ$ then the set of discontinuities of χ_E is the boundary of E, and whether or not $E \subset R^\circ$, we know that all points of $\partial(E)$ which are not contained in R° are contained in $\partial(R)$, and all discontinuities of χ_E are contained in $\partial(E)$. Hence, if E is a Jordan region then $\lambda(\partial(E)) = 0$, so the measure of the set of discontinuities of χ_E is zero, so χ_E is integrable by Theorem 12.39. If E is not a Jordan region then $\lambda(\partial(E)) \neq 0$. Since $\lambda(\partial(R)) = 0$, we know that $\lambda(\partial(E) \setminus \partial(R)) \neq 0$ since the union of measure zero sets has measure zero. Since all points of $\partial(E) \setminus \partial(R))$ are points of discontinuity of χ_E we know that the set of discontinuities of χ_E does not have Lebesgue measure zero, which means that χ_E is not integrable.

□

Theorem 12.41. *Let f be a continuous function which is bounded on $g(E)$ and let $g : E \to \mathbb{R}$ be integrable, where E is a Jordan region in $\mathbb{R}^n$. Then $f \circ g$ is integrable on E.*

Proof. If D_g is the set of discontinuities of g and $D_{f \circ g}$ is the set of discontinuities of $f \circ g$ then $D_{f \circ g} \subseteq D_g$ by Theorem 10.16. Since $\lambda(D_g) = 0$ we know that $\lambda(D_{f \circ g}) = 0$, so by the Lebesgue Characterization of Riemann Integrability it follows that $f \circ g$ is integrable on E. □

Theorem 12.42. *Let $f, g : E \to \mathbb{R}$ be integrable, where E is a Jordan region in $\mathbb{R}^n$. Then fg is integrable.*

Proof. Let D_f be the set of discontinuities of f and let D_g be the set of discontinuities of g and let D_{fg} be the set of discontinuities of fg. Then by Theorem 10.20, we know that $D_{fg} \subseteq (D_f \cup D_g)$. By the Lebesgue Characterization of Riemann Integrability, $\lambda(D_g) = \lambda(D_f) = 0$. Since the union of two sets of measure zero has measure zero and any subset of a set of measure zero has measure zero, we know that $\lambda(D_{fg}) = 0$, so fg is integrable. □

Theorem 12.43. *Let $f : E \to \mathbb{R}$ be a bounded function, where E is a Jordan region in $\mathbb{R}^n$ and R is a rectangle in $\mathbb{R}^n$ containing E. Let $g : R \to \mathbb{R}$ be a bounded function so that if $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for all $\boldsymbol{x} \in E^\circ$ and $g(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in R \setminus \overline{E}$. Then f is integrable on E if and only if g is integrable on R, in which case $\displaystyle\int_E f = \int_R g$. In particular, if H is the zero boundary extension of f to R. Then f is integrable on E if an only if H is integrable on R, in which case $\displaystyle\int_E f = \int_R H$.*

Proof. Let F be the zero extension of f to R. Recall that $E \setminus E^\circ \subseteq \overline{E} \setminus E^\circ = \partial(E)$, and $E^\circ \cup (R \setminus \overline{E}) = R \setminus \partial(E)$, which means that $F(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in R \setminus \partial(E)$. Since $\partial(E)$ is closed, if $\mathbf{x} \in R \setminus \partial(E)$ then there is a $\delta_{\mathbf{x}} > 0$ so that $B_{\delta_{\mathbf{x}}}(\mathbf{x}) \cap \partial(E) = \emptyset$.

If F is continuous at $\mathbf{p} \in R \setminus \partial(E)$ then for every $\epsilon > 0$ there is a $\delta > 0$ so that if $|\mathbf{x} - \mathbf{p}| < \delta$ and $\mathbf{x} \in R$ then $|F(\mathbf{x}) - F(\mathbf{p})| < \epsilon$. Hence, if $|\mathbf{x} - \mathbf{p}| < \min\{\delta, \delta_{\mathbf{p}}\}$ and $\mathbf{x} \in R$ then $|F(\mathbf{x}) - F(\mathbf{p})| = |g(\mathbf{x}) - g(\mathbf{p})| < \epsilon$, so g is continuous at $\mathbf{p}$. Likewise, if g is continuous on at $\mathbf{p}$ then F is continuous at $\mathbf{p}$.

Let D_F be the set of discontinuities of F and let D_g be the set of points at which g is discontinuous. Then $D_g \subseteq D_F \cup \partial(E)$. Since we know that $Vol(\partial(E)) = 0$ it follows that $\lambda(\partial(E)) = 0$. If f is integrable on E then F is integrabls on R and $\lambda(D_F) = 0$ by the Lebesgue Characterization of Riemann Integrability, which means that $\lambda(D_F \cup \partial(E)) = 0$, and so $\lambda(D_g) = 0$, which implies that g is integrable on R by the Lebesgue Characterization of Riemann Integrability. Likewise, $D_F \subseteq D_g \cup \partial(E)$, so if g is integrable on R then F is integrable on R and hence f is integrable on E.

Let $\{G_i\}$ be a sequence of grids on R so that $\{|G_i|\} \to 0$. For each grid G_k we can choose a marking T_k of G_k so that $T_k \cap \partial(E) = \emptyset$. This is because E is a Jordan region, which means that $Vol(\partial(E)) = 0$, which means that $\partial(E)$ cannot contain an n-rectangle. For each $R_j \in G_k$ it must follow that $R_j \not\subseteq \partial(E)$ so we can choose a point $t_j^* \in R_j \setminus \partial(E)$ and set T_k to be the set of points t_j^* thus chosen. Since $F(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in R \setminus \partial(E)$, we know that $\{S_{T_i}(F, G_i)\} = \{S_{T_i}(g, G_i)\}$. We know $\{S_{T_i}(F, G_i)\} \to \displaystyle\int_R F = \int_E f$ and $\{S_{T_i}(g, G_i)\} \to \displaystyle\int_R g$, which means $\displaystyle\int_E f = \int_R g$.

Finally, the zero boundary extension of f to R is just a special case of a function g as described, so the theorem follows.

□

Theorem 12.44. *Let f, g be integrable on a Jordan region E contained in a rectangle R in $\mathbb{R}^m$. Then:*

(a) Let $\alpha, \beta \in \mathbb{R}$. Then $\int_E \alpha f + \beta g = \alpha \int_E f + \beta \int_E g$.

(b) If $f(\boldsymbol{x}) \leq g(\boldsymbol{x})$ for all $\boldsymbol{x} \in E^\circ$ then $\int_E f \leq \int_E g$

(c) Let $m \in \mathbb{R}$. Then $\int_E m = m(Vol(E))$

Proof. (a) Let F, G be the zero boundary extensions of f, g to R. Then $\int_E F = \int_E f$ and $\int_E G = \int_E g$ by Theorem 12.43. Let $\{G_n\}$ be a sequence of grids whose meshes approach zero. Then by Theorem 12.30, we know that for any markings T_n of G_n it is true that $\{S_{T_n}(F, G_n)\} \to \int_R F$ and $\{S_{T_n}(G, G_n)\} \to \int_R G$. Hence, $\{S_{T_n}(\alpha F + \beta G, G_n)\} \to \alpha \int_R F + \beta \int_R G$, so, again, by Theorem 12.30, we know that $\int_E \alpha f + \beta g = \alpha \int_E f + \beta \int_E g$.

(b) By the Comparison Theorem for sequences we know that since $S_{T_n}(F, G_n) \leq S_{T_n}(G, G_n)$ for each $n \in \mathbb{N}$, it follows that $\int_E f \leq \int_E g$.

(c) Let $g(\mathbf{x}) = m$ if $\mathbf{x} \in \overline{E}$, and let $g(\mathbf{x}) = 0$ if $\mathbf{x} \in R \setminus \overline{E}$. Then by Theorem 12.43, we know that $\int_R g = \int_E m$. Let $G = \{R_i\}_{1 \leq i \leq k}$ be any grid on R. Then $U(f, G) = \sum_{R_j \in O(E,G)} m|R_j| = mV(E, G)$. Taking the infimum of both sides over all grids G on R we get $mVol(E) = \int_R g = \int_E m$.

□

Theorem 12.45. *Let E_1, E_2 be non-overlapping Jordan regions in $\mathbb{R}^m$ and let $\int_{E_1} f$, $\int_{E_2} f$ exist. Then $\int_{E_1 \cup E_1} f = \int_{E_1} f + \int_{E_2} f$.*

Proof. First, we know that f is integrable on $E_1 \cup E_2$ because by Theorem 12.39 we know that the set D_1 of discontinuities of f on E_1 and the set D_2 of discontinuities of f on E_2 have Lebesgue measure zero. Thus, the set of discontinuities of F on $E_1 \cup E_2$ is a subset of $D_1 \cup D_2 \cup \partial(E_1) \cup \partial(E_2)$. Since each of these sets has Lebesgue measure zero, we know that the set of discontinuities of F has Lebesgue measure zero, so $\int_{E_1 \cup E_2} f$ exists.

Let R be a rectangle containing $E_1 \cup E_2$. Let F, F_1, F_2 represent the zero boundary extensions of f to R considering the domain of f to be $E_1 \cup E_2$, E_1 and E_2 respectively.

Since E_1 and E_2 are non-overlapping it follows that $E_1^\circ \cap \overline{E_2} = E_2^\circ \cap \overline{E_1} = \emptyset$. Let $\{G_n\}$ be a sequence of grids on R so that $\{|G_n|\} \to 0$. Since $\partial(E_1) \cup \partial(E_2)$ has volume zero, this set contains no n-rectangles, so we can choose markings T_n for each G_n so that $T_n \cap (\partial(E_1) \cup \partial(E_2)) = \emptyset$.

For any $t_i^* \in T_n$, if $t_i^* \in E_1^\circ$ then $F_1(t_i^*) = f(t_i^*) = F(t_i^*)$ and $F_2(t_i^*) = 0$ and if $t_i^* \in E_2^\circ$ then $F_1(t_i^*) = 0$ and $F_2(t_i^*) = f(t_i^*) = F(t_i^*)$. If $t_i^* \in R$ and $t_i^* \notin E_1^\circ \cup E_2^\circ$ then $F_1(t_i^*) = 0 = F_2(t_i^*)$, and since $t_i^* \notin (\partial(E_1) \cup \partial(E_2))$ it follows that $t_i^* \notin (E_1 \cup E_2)^\circ$, so $F(t_i^*) = 0$. Thus, $S_{T_n}(F_1, G_n) + S_{T_n}(F_2, G_n) = S_{T_n}(F, G_n)$ since $F(t_i^*) = F_1(t_i^*) + F_2(t_i^*)$ for each $t_i^* \in T_i$.

Since $\{S_{T_n}(F_1, G_n)\} \to \displaystyle\int_{E_1} f$ and $\{S_{T_n}(F_2, G_n)\} \to \displaystyle\int_{E_2} f$ and $\{S_{T_n}(F, G_n)\} \to \displaystyle\int_{E_1 \cup E_2} f$, it follows that $\displaystyle\int_{E_1 \cup E_1} f = \int_{E_1} f + \int_{E_2} f$.

□

Theorem 12.46. *Let E be a Jordan region in $\mathbb{R}^n$. Then there is a $\delta > 0$ so that if G is a grid on a rectangle containing E and $|G| < \delta$ then $V(E, G) - v(E, G) < \epsilon$.*

Proof. Let $f(\mathbf{x}) = 1$ on E. By Theorem 12.29, we can find a $\delta > 0$ so that if $|G| < \delta$ then $V(E, G) - v(E, G) = U(f, G) - L(f, G) < \epsilon$. □

Definition 93

Let R be a rectangle in $\mathbb{R}^n$ containing a Jordan region E and let G be a grid on R. We define the *upper inner sum* of f with respect to grid G to be $U(f, G)^\circ = \displaystyle\sum_{R_j \in I(E,G)} M_j |R_j|$ and the *lower inner sum* of f with respect to grid G to be $L(f, G)^\circ = \displaystyle\sum_{R_j \in I(E,G)} m_j |R_j|$. We define the *upper outer sum* of f with respect to grid G to be $\overline{U}(f, G) = \displaystyle\sum_{R_j \in O(E,G)} M_j |R_j|$ and the *lower outer sum* of f with respect to grid G to be $\overline{L}(f, G) = \displaystyle\sum_{R_j \in O(E,G)} m_j |R_j|$.

Theorem 12.47. *Let $f : E \to \mathbb{R}$ be bounded, where E is a Jordan region in $\mathbb{R}^n$, and let $\epsilon > 0$. Then there is a $\delta > 0$ so that if $|G|$ is a grid on a rectangle R containing E then $U(f, G) - (U)\displaystyle\int_E f < \epsilon$ and $(L)\displaystyle\int_E f - L(f, G) < \epsilon$, and also $|U(f, G)^\circ - (U)\displaystyle\int_E f| < \epsilon$ and $|L(f, G)^\circ - (L)\displaystyle\int_E f| < \epsilon$.*

Proof. Let $W = \{W_i\}_{1 \le i \le m}$ be a grid on R so that $U(f, W) - (U)\displaystyle\int_E f < \frac{\epsilon}{3}$ and $(L)\displaystyle\int_E f - L(f, W) < \frac{\epsilon}{3}$. Choose M so that $|f(\mathbf{x})| < M$ for all $\mathbf{x} \in E$. Let $B = \displaystyle\bigcup_{i=1}^{m} \partial(W_i)$. Then $Vol(B \cup \partial(E)) = 0$.

By Theorem 12.29 we know that there is a $\delta > 0$ so that if $G = \{R_i\}_{1 \le i \le k}$ is a grid on R with $|G| < \delta$ then $-\dfrac{\epsilon}{3} < \displaystyle\sum_{R_i \in S(B \cup \partial(E), G)} -M|R_i| < \sum_{R_i \in S(B \cup \partial(E), G)} M|R_i| < \frac{\epsilon}{3}$ (using functions M and $-M$ integrated over $B \cup \partial(E)$).

We know that $U(f,G) \geq (U)\int_E f$ and $L(f,G) \leq (L)\int_E f$ by definition. We also know that $|U(f,G) - U(f,G)^\circ| = |\sum_{R_i \in S(E,G)\setminus I(E,G)} M_i|R_i|| \leq \sum_{R_i \in S(\partial(E),G)} M|R_i| < \frac{\epsilon}{3}$, and that $|L(f,G) - L(f,G)^\circ| = |\sum_{R_i \in S(E,G)\setminus I(E,G)} m_i|R_i|| \leq \sum_{R_i \in S(\partial(E),G)} M|R_i| < \frac{\epsilon}{3}$.

We also know that if $R_i \subseteq W_j^\circ$ then $M_j^W \geq M_i^G \geq m_i^G \geq m_j^W$. Thus, $\sum_{R_i \subset W_j^\circ} M_i^G|R_i| \leq M_j|W_j|$ and $\sum_{R_i \subset W_j^\circ} m_i^G|R_i| \geq m_j \sum_{R_i \subset W_j^\circ} |R_i|$ for each $1 \leq j \leq m$.

From this we conclude that $U(f,G) = \sum_{j=1}^m \sum_{R_i \subset W_j^\circ} M_i^G|R_i| + \sum_{R_i \in S(B,G)} M_i^G|R_i| \leq \sum_{j=1}^m M_j|W_j| + \sum_{R_i \in S(B,G)} M|R_i| < U(f,W) + \frac{\epsilon}{3} < (U)\int_E f + \frac{2\epsilon}{3}$. Hence, we see that $U(f,G) - (U)\int_E f < \frac{2\epsilon}{3}$ and $|U(f,G)^\circ - (U)\int_E f| < \frac{2\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$.

Similarly, $L(f,G) = \sum_{j=1}^m \sum_{R_i \subset W_j^\circ} m_i^G|R_i| + \sum_{R_i \in S(B,G)} m_i^G|R_i|$. Now, $L(f,W) = \sum_{i=1}^m m_j^W|W_i| = \sum_{j=1}^m \sum_{R_i \subset W_j^\circ} m_j^W|R_i| + \sum_{i=1}^m \sum_{R_i \in S(B,G)} m_j^W|R_i \cap W_j|$. We know that $|\sum_{i=1}^m \sum_{R_i \in S(B,G)} m_j^W|R_i \cap W_j|| \leq \sum_{i=1}^m \sum_{R_i \in S(B,G)} M|R_i \cap W_j| = \sum_{R_i \in S(B,G)} M|R_i| < \frac{\epsilon}{3}$. Since $\sum_{j=1}^m \sum_{R_i \subset W_j^\circ} m_i^G|R_i| \geq \sum_{j=1}^m \sum_{R_i \subset W_j^\circ} m_j^W|R_i|$, we conclude that $L(f,G) \geq L(f,W) - \frac{\epsilon}{3}$, so $(L)\int_E f - L(f,G) < \frac{2\epsilon}{3}$ and $|L(f,G)^\circ - \int_E f| < \epsilon$.

□

Theorem 12.48. *Let E be a Jordan region in $\mathbb{R}^n$ and let $G = \{R_i\}_{1 \leq i \leq k}$ be a grid on a rectangle containing E and let f be a bounded function on E with $|f(\boldsymbol{x})| < M$ for all $\boldsymbol{x} \in E$. Then $|\overline{U}(f,G) - U(f,G)^\circ| \leq MV(\partial(E),G)$, $|\overline{U}(f,G) - U(f,G)| \leq MV(\partial(E),G)$, $|\overline{L}(f,G) - L(f,G)| \leq MV(\partial(E),G)$ and $|\overline{L}(f,G) - L(f,G)^\circ| \leq MV(\partial(E),G)$.*

Proof. By definition $|\overline{U}(f,G) - U(f,G)^\circ| = |\sum_{R_i \in O(E,G)} M_i|R_i| - \sum_{R_i \in I(E,G)} M_i|R_i|| = |\sum_{R_i \in S(\partial(E),G)} M_i|R_i|| \leq MV(\partial(E),G)$. Likewise, since $O(E,G) \setminus I(E,G) \subseteq O(E,G) \setminus S(E,G)$ we know $|\overline{U}(f,G) - U(f,G)| = |\sum_{R_i \in O(E,G)} M_i|R_i| - \sum_{R_i \in S(E,G)} M_i|R_i|| \leq MV(\partial(E),G)$.

Similarly, $|\overline{L}(f,G) - L(f,G)^\circ| = |\sum_{R_i \in O(E,G)} m_i|R_i| - \sum_{R_i \in I(E,G)} m_i|R_i|| = |\sum_{R_i \in S(\partial(E),G)} m_i|R_i|| \leq MV(\partial(E),G)$ and $|\overline{L}(f,G) - L(f,G)| = |\sum_{R_i \in O(E,G)} m_i|R_i| - \sum_{R_i \in S(E,G)} m_i|R_i|| \leq MV(\partial(E),G)$.

□

Theorem 12.49. *Let $E \subseteq \mathbb{R}^n$ be a Jordan region contained in a rectangle R and let $f : R \to \mathbb{R}$ be bounded. Then the following are equivalent:*

(a) f is integrable on E

(b) For every $\epsilon > 0$ there is a $\delta > 0$ so that if G is a grid on a rectangle containing E with $|G| < \delta$ then all upper and lower sums, upper and lower outer sums and upper and lower inner sums are within a distance ϵ of each other.

(c) For every $\epsilon > 0$ there is a $\delta > 0$ so that if G is a grid on a rectangle containing E with $|G| < \delta$ then one of the upper sums listed (upper, upper inner or upper outer) is within a distance ϵ of one of the lower sums listed (upper, upper inner or upper outer).

Proof. Let $\epsilon > 0$. Choose $M > 0$ so that $|f(\mathbf{x})| \leq M$ on R. Since E is a Jordan region we can find a $\delta_1 > 0$ so that if $|G| < \delta_1$ then $V(\partial(E), G) < \dfrac{\epsilon}{4M}$.

(a) implies (b). By Theorem 12.47 we can find $0 < \delta < \delta_1$ so that if $|G| < \delta_1$ then $|U(f,G)^\circ - \displaystyle\int_E f| < \frac{\epsilon}{2}$ and $|L(f,G)^\circ - \displaystyle\int_E f| < \frac{\epsilon}{2}$ and $|U(f,G) - \displaystyle\int_E f| < \frac{\epsilon}{2}$ and $|L(f,G) - \displaystyle\int_E f| < \frac{\epsilon}{2}$. By Theorem 12.48, we know that $|\overline{U}(f,G) - U(f,G)^\circ| < M\dfrac{\epsilon}{4M} < \dfrac{\epsilon}{2}$. The result follows from the triangle inequality.

(b) implies (c). This is immediate.

(c) implies (a). By Theorem 12.47, we know that we can find $\delta < \delta_1$ so that the upper sums, upper outer sums, and upper inner sums over a grid G with $|G| < \delta$ are within distance $\dfrac{\epsilon}{4}$ of one another, and the lower inner sums, lower outer sums and lower sums are also within a distance $\dfrac{\epsilon}{4}$ of one another. Hence, if any of these upper sums can be made within distance $\dfrac{\epsilon}{2}$ of any of the lower sums then by the triangle inequality it follows that $U(f,G) - L(f,G) < \epsilon$ so f is integrable. □

Iterated Integrals

It is frequently possible to express integrals over a Jordan region or an rectangle in $\mathbb{R}^n$ as an iterated integral, where one integrates integrals with respect to previous variables, applying the Fundamental Theorem of Calculus each time in order to end up with a neat way to evaluate an integral.

Definition 94

A set $E \subset \mathbb{R}^2$ is a *type one region* if $E = \{(x,y) \in \mathbb{R}^2 | a \leq x \leq b \text{ and } g_1(x) \leq y \leq g_2(x)\}$, where g_1, g_2 are continuous functions of x on $[a,b]$. We say $E \subset \mathbb{R}^2$ is a *type two region* if $E = \{(x,y) \in \mathbb{R}^2 | c \leq y \leq d \text{ and } g_1(y) \leq x \leq g_2(y)\}$, where g_1, g_2 are continuous functions of y on $[c,d]$. The functions g_1, g_2 and will be referred to as *boundary functions* for the region in both definitions.

Let D be a region in the plane that is a Jordan region. A set $E \subset \mathbb{R}^3$ is a *type* *one region* if $E = \{(x, y, z)|(x, y) \in D$ and $g_1(x, y) \leq z \leq g_2(x, y)\}$, where g_1, g_2 are continuous functions on D, a set in the plane. We say E is a *type two region* if $E = \{(x, y, z)|(y, z) \in D$ and $g_1(y, z) \leq x \leq g_2(y, z)\}$, where g_1, g_2 are continuous functions on D. We say E is a *type three region* if $E = \{(x, y, z)|(x, z) \in D$ and $g_1(x, z) \leq y \leq g_2(x, z)\}$, where g_1, g_2 are continuous functions on D referred to as boundary functions for these regions.

More generally, we can define a *projectable* region in $\mathbb{R}^n$ to be a region $E = \{(\mathbf{x}, x_j, \mathbf{y}) \in \mathbb{R}^n|(\mathbf{x}, \mathbf{y}) \in D$ and $g_1((\mathbf{x}, \mathbf{y})) \leq x_n \leq g_2((\mathbf{x}, \mathbf{y}))$ for all $(\mathbf{x}, \mathbf{y}) \in D\}$, where D is a closed Jordan region in $\mathbb{R}^{n-1}$ and $g_1 < g_2$ and g_1, g_2 are continuous on D. We can inductively define E to be *fully projectable* if D is fully projectable.

The theorems that follow justify when we can use iterated integration, which is a process we will then describe.

Theorem 12.50. *Let D be a closed Jordan region in $\mathbb{R}^n$, and let $f, g : D \to \mathbb{R}$ be continuous functions so that $f(x) \leq g(x)$ on D. Let E be the projectable region $\{(\boldsymbol{x}, x_{n+1}) \in \mathbb{R}^{n+1}|\boldsymbol{x} \in D$ and $f(\boldsymbol{x}) \leq x_{n+1} \leq g(\boldsymbol{x})\}$ in $\mathbb{R}^{n+1}$. Let $G_1 = \{(\boldsymbol{x}, f(\boldsymbol{x})) \in \mathbb{R}^{n+1}|\boldsymbol{x} \in D\}$ and $G_2 = \{(\boldsymbol{x}, g(\boldsymbol{x})) \in \mathbb{R}^{n+1}|\boldsymbol{x} \in D\}$ and $W = \{(\boldsymbol{x}, x_{n+1})|\boldsymbol{x} \in \partial(D)$ and $f(\boldsymbol{x}) \leq x_{n+1} \leq g(\boldsymbol{x})\}$ then $\partial(E) = G_1 \cup G_2 \cup W$. Then E is a compact Jordan region and $\partial(E) = G_1 \cup G_2 \cup W$.*

Proof. Let $\mathbf{p} \in \partial(E)$ and let $\mathbf{q} = (p_1, p_2, p_3, ..., p_n)$ so $\mathbf{p} = (\mathbf{q}, p_{n+1})$. If $\mathbf{q} \in D^\circ$ then there is an open ball $B_{\epsilon_1}(\mathbf{q})$ contained in D. Suppose $f(\mathbf{q}) < p_{n+1} < g(\mathbf{q})$. Let $\gamma = \min\{p_{n+1} - f(\mathbf{q}), g(\mathbf{q}) - p_{n+1}\}$. Since f and g are continuous we can find $0 < \delta < \epsilon_1$ so that if $|\mathbf{x} - \mathbf{q}| < \delta$ in $\mathbb{R}^n$ then $|f(\mathbf{x}) - f(\mathbf{q})| < \frac{\gamma}{2}$. Let $\epsilon_2 = \min\{\delta, \frac{\gamma}{2}\}$. Then $B_{\epsilon_2}(\mathbf{p}) \subseteq E$ because any point $(\mathbf{x}, x_{n+1}) \in B_{\epsilon_2}(\mathbf{p})$ is a point where $|\mathbf{x} - \mathbf{q}| < \delta$, so $f(\mathbf{x}) < f(\mathbf{q}) - \frac{\gamma}{2} < x_{n+1} < g(\mathbf{q}) + \frac{\gamma}{2} < g(\mathbf{q})$. This contradicts $\mathbf{p}$ being a boundary point of E. We conclude that either $p_{n+1} = f(\mathbf{q})$ or $p_{n+1} = g(\mathbf{q})$.

If $\mathbf{q} \notin \overline{D}$ then for some $\epsilon_3 > 0$ we know that $B_{\epsilon_3}(\mathbf{q}) \cap \overline{D} = \emptyset$ in $\mathbb{R}^n$. For every $(\mathbf{x}, t) \in E$ we know that $\mathbf{x} \in D$, so it follows that $B_{\epsilon_3}(\mathbf{p}) \cap \overline{D} = \emptyset$ in $\mathbb{R}^{n+1}$, so $\mathbf{p} \notin \partial(E)$. Hence, if $\mathbf{p} \notin G_1 \cup G_2$ then $\mathbf{q} \in \overline{D} \setminus D^\circ = \partial(D)$. Thus, $\partial(E) \subseteq G_1 \cup G_2 \cup W$.

Next, let $\mathbf{z} \in G_1 \cup G_2 \cup W$, where $\mathbf{z} = (\mathbf{y}, z_{n+1})$ for $\mathbf{y} = (z_1, z_2, z_3, ..., z_n) \in \mathbb{R}^n$. If $\mathbf{z} \in G_1$ then $z_{n+1} = f(\mathbf{y})$ and so for any $\epsilon > 0$ it follows that $B_\epsilon(\mathbf{z})$ contains both $\mathbf{z} \in E$ and $(\mathbf{y}, f(\mathbf{y}) - \frac{\epsilon}{2}) \notin E$, which means that $\mathbf{z} \in \partial(E)$. Likewise, if $\mathbf{z} \in G_2$ then $z_{n+1} = g(\mathbf{y})$ and so for any $\epsilon > 0$ it follows that $B_\epsilon(\mathbf{z})$ contains both $\mathbf{z} \in E$ and $(\mathbf{y}, g(\mathbf{y}) + \frac{\epsilon}{2}) \notin E$, which means that $\mathbf{z} \in \partial(E)$.

Let $\mathbf{z} \in W \setminus (G_1 \cup G_2)$. Then for any $\epsilon > 0$ we know that $B_\epsilon(\mathbf{y})$ (in $\mathbb{R}^n$) contains a point $\mathbf{s} \notin D$, and hence $B_\epsilon(\mathbf{z})$ (in $\mathbb{R}^{n+1}$) contains the point $(\mathbf{s}, z_{n+1}) \notin E$ as well as the point $\mathbf{z} \in E$, which means that $\mathbf{z} \in \partial(E)$. Hence, $\partial(E) = G_1 \cup G_2 \cup W$.

Since the boundary of E is contained in E we know that E is closed, so by the Heine-Borel Theorem E is compact.

To show that E is a Jordan region, we will show $Vol(G_1) = Vol(G_2) = Vol(W) = 0$. Let $\epsilon > 0$. Let R be a rectangle containing E in $\mathbb{R}^n$. Since E is compact, f and g are uniformly continuous on E. Choose $\delta > 0$ so that if $\mathbf{x}, \mathbf{y} \in E$ and $|\mathbf{x} - \mathbf{y}| < \delta$ then

$|f(\mathbf{x})-f(\mathbf{y})|<\frac{\epsilon}{2|R|}$. Choose a grid $G=\{R_i\}_{1\leq i\leq k}$ on R so that $|G|<\delta$. For each $R_i\in G$ so that $R_i\cap E\neq\emptyset$, choose a point $\mathbf{x}_i\in R_i$ and let $Q_i=R_i\times[f(\mathbf{x}_i)-\frac{\epsilon}{2|R|},f(\mathbf{x}_i)+\frac{\epsilon}{2|R|}]$. Then $\{(\mathbf{x},f(\mathbf{x}))\in G_1|\mathbf{x}\in R_i\}\subseteq Q_i$. Thus, $W_1=\{Q_i|R_i\cap E\neq\emptyset\}$ is a set of rectangles that covers G_1 and $\displaystyle\sum_{Q_i\in W_1}|Q_i|\leq\sum_{i=1}^{k}|R_i|\frac{\epsilon}{|R|}\leq\epsilon$. Hence, $Vol(G_1)=0$. Similarly, replacing f by g and G_1 by G_2 we see that $Vol(G_2)=0$.

Since E is compact, f and g are bounded on E by the Extreme Value Theorem, so we pick m,M so that $m\leq f(\mathbf{x})\leq g(\mathbf{x})<M$ for all $\mathbf{x}\in D$. Since $\partial(D)$ has volume zero we can find a collection of rectangles $\{K_i\}_{1\leq i\leq t}$ covering D in $\mathbb{R}^n$ so that $\displaystyle\sum_{i=1}^{t}|K_i|<\frac{\epsilon}{M-m}$. For each $1\leq i\leq t$, define $S_i=K_i\times[m,M]$. Then $\{S_i\}_{1\leq i\leq t}$ is a collection of rectangles that covers W and $\displaystyle\sum_{i=1}^{t}|S_i|<(M-m)\frac{\epsilon}{M-m}=\epsilon$. Hence $Vol(W)=0$. It follows that $Vol(\partial(E))=0$, so E is a Jordan region.

□

What follows is a generalization of Fubini's Theorem in two variables (though it is not as strong as what is usually thought of as Fubini's Theorem in $\mathbb{R}^n$). Fubini's original theorem in two variables is part (c).

Theorem 12.51. *Fubini's Theorem. Let $f:R=[a_1,b_1]\times[a_2,b_2]\to\mathbb{R}$ be an integrable function on R.*

(a) Let $\displaystyle\int_{a_2}^{b_2}f(x,y)dy$ exist for each $x\in[a_1,b_1]$. Then $\displaystyle\int_R f=\int_{a_1}^{b_1}(\int_{a_2}^{b_2}f(x,y)dy)dx$.

(b) Let $\displaystyle\int_{a_1}^{b_1}f(x,y)dx$ exist for each $y\in[a_2,b_2]$. Then $\displaystyle\int_R f=\int_{a_2}^{b_2}(\int_{a_1}^{b_1}f(x,y)dy)dx$.

(c) Let $\displaystyle\int_{a_2}^{b_2}f(x,y)dy$ exist for each $x\in[a_1,b_1]$ and let $\displaystyle\int_{a_1}^{b_1}f(x,y)dx$ exist for each $y\in[a_2,b_2]$. Then $\displaystyle\int_R f=\int_{a_2}^{b_2}(\int_{a_1}^{b_1}f(x,y)dy)dx=\int_{a_1}^{b_1}(\int_{a_2}^{b_2}f(x,y)dy)dx$.

(d) Let $g:Q\to\mathbb{R}$ be integrable, where $Q=\displaystyle\prod_{i=1}^{n}[a_i,b_i]$ is a rectangle in $\mathbb{R}^n$, and $D=\displaystyle\prod_{i=1}^{n-1}[a_i,b_i]$ is a rectangle in $\mathbb{R}^{n-1}$. For each $\boldsymbol{x}\in\displaystyle\prod_{i=1}^{n-1}[a_i,b_i]=Q_{n-1}$, let $g(\boldsymbol{x},t):[a_n,b_n]\to\mathbb{R}$ also be integrable. Then $\displaystyle\int_Q g=\int_D\int_{a_n}^{b_n}g(\boldsymbol{x},t)$.

(e) Let E_n be a fully projectable Jordan region contained in Q so that with respect to a particular ordering of the variables $x_1,x_2,...,x_n$ (after a possible re-labeling), so that there are Jordan regions $E_1,E_2,...,E_{n-1}$ so that $E_k\subseteq\displaystyle\prod_{i=1}^{k}[a_i,b_i]=Q_k\subset\mathbb{R}^k$ for each $k\in\{2,3,...,n-1\}$ and $E_1=[a_1,b_1]$, and also continuous functions $f_k,g_k:E_k\to\mathbb{R}$ for

each $k \in \{1,2,3,...,n-1\}$ *so that* $E_{k+1} = \{(\boldsymbol{x},t) \in \mathbb{R}^{k+1} | \boldsymbol{x} \in E_k \textit{ and } f_k(\boldsymbol{x}) \le t \le g_k(\boldsymbol{x})\ \}$ *for all* $k \in \{1,2,...,n-1\}$*. Then*

$$\int_E g = \int_{a_1}^{b_1} \int_{f_1(x_1)}^{g_1(x_1)} \int_{f_2(x_1,x_2)}^{g_2(x_1,x_2)} ... \int_{f_{n-1}(x_1,...,x_{n-1})}^{g_{n-1}(x_1,...,x_{n-1})} g(x_1,x_2,...,x_n)dx_ndx_{n-1}...dx_1$$

.

Proof. (a) Let G be the grid induced by partitions $P_1 = \{x_0, x_1, ..., x_n\}$ of $[a_1, b_1]$ and $P_2 = \{y_0, y_1, ..., y_m\}$ of $[a_2, b_2]$. Let M_{ij}, m_{ij} denote the supremum and infimum of $f(x,y)$ over $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$, for each $1 \le i \le n$ and $1 \le j \le m$. For a given integer $i \in [1,n]$, if $x_i^* \in [x_{i-1}, x_i]$ and $y \in [y_{j-1}, y_j]$ it is true that $m_{ij} \le f(x_i^*, y) \le M_{ij}$. Then $m_{ij}(y_j - y_{j-1}) \le \int_{y_{j-1}}^{y_j} f(x_i^*, y)dy \le M_{ij}(y_j - y_{j-1})$. Thus, it follows that $\sum_{j=1}^{m} m_{ij}(y_j - y_{j-1}) \le \sum_{j=1}^{m} \int_{y_{j-1}}^{y_j} f(x_i^*, y)dy = \int_{a_2}^{b_2} f(x_i^*, y)dy \le \sum_{j=1}^{m} M_{ij}(y_j - y_{j-1})$. Taking the sum over $1 \le i \le n$ gives us $\sum_{i=1}^{n}\sum_{j=1}^{m} m_{ij}(x_i - x_{i-1})(y_j - y_{j-1}) \le \sum_{i=1}^{n} \int_{a_2}^{b_2} f(x_i^*, y)dy(x_i - x_{i-1}) \le \sum_{i=1}^{n}\sum_{j=1}^{m} M_{ij}(x_i - x_{i-1})(y_j - y_{j-1})$. Setting $M_i = \sup_{x_i^* \in [x_{i-1}, x_i]} \int_{a_2}^{b_2} f(x_i^*, y)dy$ and $m_i = \inf_{x_i^* \in [x_{i-1}, x_i]} \int_{a_2}^{b_2} f(x_i^*, y)dy$, we see that $\sum_{i=1}^{n}\sum_{j=1}^{m} m_{ij}(x_i - x_{i-1})(y_j - y_{j-1}) \le \sum_{i=1}^{n} m_i(x_i - x_{i-1}) \le \sum_{i=1}^{n} M_i(x_i - x_{i-1}) \le \sum_{i=1}^{n}\sum_{j=1}^{m} M_{ij}(x_i - x_{i-1})(y_j - y_{j-1})$. In other words, $L(f,G) \le L(\int_{a_2}^{b_2} f(x_i^*, y), P_1) \le U(\int_{a_2}^{b_2} f(x_i^*, y), P_1) \le U(f,G)$. We can, for every $\epsilon > 0$, find a grid G so that $U(f,G) - L(f,G) < \epsilon$, which means that we can make $\sum_{i=1}^{n} M_i(x_i - x_{i-1}) - \sum_{i=1}^{n} m_i(x_i - x_{i-1}) < \epsilon$, from which we conclude that $\int_{a_1}^{b_1}(\int_{a_2}^{b_2} f(x,y)dy)dx$ exists, and since it is less than or equal to all upper sums and greater than or equal to all lower sums for f, we know $\int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x,y)dydx = \int_R f$.

(b) This follows from (a) by switching the labels of the variables.

(c) This is an immediate consequence of (a) and (b).

(d) This is similar to the proof of (a). Let $\epsilon > 0$. Since g is integrable on Q we can find a grid $G = \{Q_i\}_{1 \le i \le t}$ on Q induced by the set of partitions $P = \{P_1, P_2, ..., P_n\}$, where P_i is a partition of $[a_i, b_i]$ for each $1 \le i \le n$, so that $U(g,G) - L(g,G) < \epsilon$. Let $H = \{D_i\}_{1 \le i \le s}$ be the grid on Q_{n-1} induced by $P_1, P_2, ..., P_{n-1}$.

Let $P_n = \{z_1, z_2, ..., z_m\}$. Let M_{ij}, m_{ij} denote the supremum and infimum respectively of $g(x,y)$ over $D_i \times [z_{j-1}, z_j]$ for each $1 \le i \le s$ and $1 \le j \le m$. Then for any $D_i \in H$ if we pick $\mathbf{x}_i^* \in D_i$ it follows that $m_{ij} \le g(\mathbf{x}_i^*, t) \le M_{ij}$ if $z_{j-1} \le t \le z_j$.

Hence, it follows that $m_{ij}(z_j - z_{j-1}) \le \int_{z_{j-1}}^{z_j} g(\mathbf{x}_i^*, t)dt \le M_{ij}(z_j - z_{j-1})$. Thus,

$$\sum_{j=1}^{m} m_{ij}(z_i - z_{i-1}) \leq \sum_{j=1}^{m} \int_{z_{i-1}}^{z_i} g(\mathbf{x}_i^*, t)dt = \int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt \leq \sum_{j=1}^{m} M_{ij}(z_i - z_{i-1}).$$

From this, we see that $L(g, G) = \sum_{D_i \in H} \sum_{j=1}^{m} m_{ij}|D_i|(z_j - z_{j-1}) \leq \sum_{D_i \in H} |D_i| \int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt \leq \sum_{D_i \in H} \sum_{j=1}^{m} M_{ij}|D_i|(z_j - z_{j-1}) = U(g, G)$.

Setting $M_i = \sup_{\mathbf{x}_i^* \in D_i} \int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt$ and $m_i = \inf_{\mathbf{x}_i^* \in D_i} \int_{a_n}^{b^n} g(\mathbf{x}_i^*, t)dt$, the statement in the preceding paragraph gives us that $L(g, G) \leq \sum_{D_i \in H} m_i|D_i| \leq \sum_{D_i \in H} M_i|D_i| \leq U(g, G)$. Since $\sum_{D_i \in H} m_i|D_i| = L(\int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt, H)$ and $\sum_{D_i \in H} M_i|D_i| = U(\int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt, H)$, this tells us that $U(\int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt, H) - L(\int_{a_n}^{b_n} g(\mathbf{x}_i^*, t)dt, H) \leq U(g, G) - L(g, G) < \epsilon$, which means that $\int_D \int_{a_n}^{b_n} g(\mathbf{x}, t)dt$ exists. Furthermore, since $\int_Q g, \int_D \int_{a_n}^{b_n} g(\mathbf{x}, t)dt \in [L(g, G), U(g, G)]$ we know that $|\int_Q g - \int_D \int_{a_n}^{b_n} g(\mathbf{x}, t)dt| < \epsilon$. Since this is true for all $\epsilon > 0$, we conclude that $\int_D \int_{a_n}^{b_n} g(\mathbf{x}, t)dt = \int_Q g$.

(e) This follows inductively from (d) and the fact that the integral of a function over a Jordan region and the integral of its zero extension are the same. Since we are integrating over E, we re-define g to be the zero extension of the original function g, extending g to Q, and note that $\int_Q g = \int_E g$.

By (b) we have that $\int_Q g = \int_{Q_{n-1}} \int_{a_n}^{b_n} g(\mathbf{x}, x_n)dx_n$, where $\mathbf{x} \in Q_{n-1}$. Since the only non-zero values of $g(\mathbf{x}, x_n)$ occur for $f_{n-1}(\mathbf{x}) \leq x_n \leq g_{n-1}(\mathbf{x})$ it follows that $\int_E g = \int_{Q_{n-1}} \int_{f_{n-1}(\mathbf{x})}^{g_{n-1}(\mathbf{x})} g(\mathbf{x}, x_n)dx_n$. Then, treating $\int_{f_{n-1}(\mathbf{x})}^{g_{n-1}(\mathbf{x})} g(\mathbf{x}, x_n)dx_n$ as the integrand function, we apply (b) again to give us that
$\int_E g = \int_{Q_{n-2}} \int_{a_{n-1}}^{b_{n-1}} \int_{f_{n-1}(\mathbf{x}, x_{n-1})}^{g_{n-1}(\mathbf{x}, x_{n-1})} g(\mathbf{x}, x_{n-1}, x_n)dx_n dx_{n-1}$, where $\mathbf{x} \in Q_{n-2}$. Since we know that if $x_{n-1} \notin [f_{n-2}(\mathbf{x}), g_{n-2}(\mathbf{x})]$ then $g(\mathbf{x}, x_{n-1}, x_n) = 0$ (since $(\mathbf{x}, x_{n-1}, x_n) \notin E$), it follows that $\int_E g = \int_{Q_{n-2}} \int_{f_{n-1}(\mathbf{x})}^{g_{n-1}(\mathbf{x})} \int_{f_{n-1}(\mathbf{x}, x_{n-1})}^{g_{n-1}(\mathbf{x}, x_{n-1})} g(\mathbf{x}, x_{n-1}, x_n)dx_n dx_{n-1}$. Repeating this process yields the indicated formula that

$$\int_E g = \int_{a_1}^{b_1} \int_{f_1(x_1)}^{g_1(x_1)} \int_{f_2(x_1, x_2)}^{g_2(x_1, x_2)} \ldots \int_{f_{n-1}(x_1, \ldots, x_{n-1})}^{g_{n-1}(x_1, \ldots, x_{n-1})} g(x_1, x_2, \ldots, x_n)dx_n dx_{n-1} \ldots dx_1$$

.

□

This theorem doesn't just tell us that we can switch the order of integration. It also tells us that we can represent integrals as iterated integrals. Integral order can be rearranged into any order of variables for which the Jordan region is fully projectable in the manner described in part (e).

Fubini's Theorem also lets us prove Clairaut's Theorem (already proven) more simply, and since its proof did not depend on Clairaut's Theorem there is some value to giving this simpler proof here.

Theorem 12.52. *Clairaut's Theorem (again). Let $f : V \to \mathbb{R}$ be C^2, where V is an open set in $\mathbb{R}^2$. Then $f_{xy} = f_{yx}$ on V.*

Proof. Let $(x_0, y_0) \in V$. Since V is open we can choose $\Delta x, \Delta y$ small enough so that the rectangle $R = [x_0, x_0 + \Delta x] \times [y_0, y_0 + \Delta y] \subset V$. By the Fundamental Theorem of Calculus, $\int_{x_0}^{x_0+\Delta x} \int_{y_0}^{y_0+\Delta y} f_{xy}(x,y)dydx = \int_{x_0}^{x_0+\Delta x} f_x(x, y_0 + \Delta y) - f_x(x, y_0)dx = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 + \Delta x, y_0) - f(x_0, y_0 + \Delta y) + f(x_0, y_0)$. By Fubini's Theorem, it follows that $f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 + \Delta x, y_0) - f(x_0, y_0 + \Delta y) + f(x_0, y_0) = \int_R f_{xy}$.

Likewise, if we were to integrate f_{yx} we would get $\int_{y_0}^{y_0+\Delta y} \int_{x_0}^{x_0+\Delta x} f_{yx}(x,y)dxdy = \int_{y_0}^{y_0+\Delta y} f_y(x_0 + \Delta x, y) - f_y(x_0, y)dy = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 + \Delta x, y_0) - f(x_0, y_0 + \Delta y) + f(x_0, y_0) = \int_R f_{yx}$. Thus, $\int_R f_{xy} - f_{yx} = 0$.

Suppose that $|f_{xy}(x_0, y_0) - f_{yx}(x_0, y_0)| = \epsilon > 0$. Then since $f_{xy} - f_{yx}$ is continuous we can find a $\delta > 0$ so that if $|(x,y) - (x_0, y_0)| < \delta$ then $|(f_{xy} - f_{yx})(x,y) - (f_{xy} - f_{yx})(x_0, y_0)| < \frac{\epsilon}{2}$ which means that $|f_{xy}(x_0, y_0) - f_{yx}(x_0, y_0)| > \frac{\epsilon}{2}$. Choosing $\Delta x, \Delta y < \frac{\delta}{\sqrt{2}}$ it must follow that $|f_{xy}(x,y) - f_{yx}(x,y)| > \frac{\epsilon}{2}$ and $f_{xy}(x,y) - f_{yx}(x,y)$ is either positive on $|R|$ or negative on R, which means that $|\int\int_R f_{xy}(x,y) - f_{yx}(x,y)dA| \geq \Delta x \Delta y (\frac{\epsilon}{2})$, a contradiction.

It follows that $f_{xy}(x_0, y_0) = f_{yx}(x_0, y_0)$. □

The ideas of area and volume now have multiple definitions in terms of integrals. Recall, for instance, that the area between two functions was supposed to be the integral of the difference of these functions, $\int_a^b g(x) - f(x)dx$. However, we also have a version of area (or two-volume) defined by inner and outer sums which is $\int_R 1dA$, which by the previous theorem is $\int_a^b \int_{f(x)}^{g(x)} 1dA$, which is $\int_a^b g(x) - f(x)dx$, so the two notions of are the same. While this type of observation lets us determine that integrals give volumes which agree with the preceding section's definition of volume, we have to work slightly harder in order to get that the notions of volume are the same for regions that are not fully projectable. This is the objective of the theorem below.

Theorem 12.53. *Let D be a closed Jordan region in $\mathbb{R}^n$, and let $f, g : D \to \mathbb{R}$ be continuous functions so that $f(x) \leq g(x)$ on D. Then $Vol(E) = \int_E g - f$.*

Proof. Let $G_1 = \{(\mathbf{x}, f(\mathbf{x})) \in \mathbb{R}^{n+1} | \mathbf{x} \in D\}$ and $G_2 = \{(\mathbf{x}, g(\mathbf{x})) \in \mathbb{R}^{n+1} | \mathbf{x} \in D\}$ and $W = \{(\mathbf{x}, x_{n+1}) | \mathbf{x} \in \partial(D) \text{ and } f(\mathbf{x}) \leq x_{n+1} \leq g(\mathbf{x})\}$ then $\partial(E) = G_1 \cup G_2 \cup W$. Recall that by Theorem 12.50, E is a compact Jordan region and $\partial(E) = G_1 \cup G_2 \cup W$.

Let $\epsilon > 0$. We first show the result is true if $f(\mathbf{x}) < g(\mathbf{x})$ on D. Let $m = \min_D g(\mathbf{x}) - f(\mathbf{x}) > 0$. Since f and g are uniformly continuous on D we can find $\delta > 0$ so that if $|\mathbf{x} - \mathbf{y}| < \delta$ then $|g(\mathbf{x}) - g(\mathbf{y})| < \frac{m}{2}$ and $|f(\mathbf{x}) - f(\mathbf{y})| < \frac{m}{2}$ and therefore $g(\mathbf{x}) > f(\mathbf{y})$.

By Theorem 12.49 we can find a grid $G = \{R_i\}_{1 \leq i \leq k}$ on R on a rectangle containing D so that $|G| < \delta$ and all upper and lower sums and upper and lower inner and outer sums of f and g with respect to G are within $\frac{\epsilon}{2}$ of the respective functions f and g. In particular, $\sum_{R_i \in I(D,G)} (m_i(g) - M_i(f))|R_i| - (\int_D g - \int_D f) < \epsilon$. The rectangles $R_i \times [M_i(f), m_i(g)]$ are contained in the interior of E if $R_i \in I(D, G)$ because $f(\mathbf{x}) < g(\mathbf{x})$ on each R_i. Hence, $Vol(E) \geq \sum_{R_i \in I(D,G)} (m_i(g) - M_i(f))|R_i|$. Likewise, $\sum_{R_i \in O(D,G)} (M_i(g) - m_i(f))|R_i| - (\int_D g - \int_D f) < \epsilon$, and $E \subseteq \bigcup_{R_i \in O(D,G)} R_i \times [m_i(f), M_i(g)]$ because each point of E has coordinates other than the x_{n+1} coordinate which are inside D and if $(\mathbf{x}, t) \in E$ then $\mathbf{x} \in R_i$ for some $R_i \in O(E, G)$ and $t \in [m_i(f), M_i(G)]$. Thus, $Vol(E) \leq \sum_{R_i \in O(D,G)} (M_i(g) - m_i(f))|R_i|$. Since both $\sum_{R_i \in O(D,G)} (M_i(g) - m_i(f))|R_i|$ and $\sum_{R_i \in O(D,G)} (M_i(g) - m_i(f))|R_i|$ are within a distance ϵ of $\int_D g - \int_D f$ we conclude that $|Vol(E) - \int_D g - f| < \epsilon$ for all $\epsilon > 0$ and therefore $Vol(E) = \int_D g - f$.

Next, let $f(\mathbf{x}) \leq g(\mathbf{x})$ on D and extend f and g to their zero extensions on R. Choose a $\delta > 0$ so that if $|\mathbf{x} - \mathbf{y}| < \delta$ then $|f(\mathbf{x}) - f(\mathbf{y})| < \frac{\epsilon}{2|R|}$ and $|g(\mathbf{x}) - g(\mathbf{y})| < \frac{\epsilon}{2|R|}$.

Choose grid $G = \{R_i\}_{1 \leq i \leq k}$ on R with $|G| < \delta$.

Let $W = \{R_i \in G | m_i(g) - M_i(f) > 0\}$. For each $R_i \in W$ we know that $Vol((R_i \times \mathbb{R}) \cap E) = \int_{R_i} g - f$ by the preceding argument.

For each $R_i \in G \setminus W$ we know that $(R_i \times \mathbb{R}) \cap E \subseteq R_i \times [m_i(f), M_i(g)]$, and that $M_i(g) - m_i(f) \leq M_i(g) - m_i(g) + M_i(f) - m_i(f) < \frac{\epsilon}{|R|}$ since $M_i(f) \geq m_i(g)$. Thus, $Vol(E \setminus \bigcup_{R_i \in W} Vol((R_i \times \mathbb{R}) \cap E)) < \sum_{R_i \in G \setminus W} |R_i| \frac{\epsilon}{|R|} \leq \epsilon$. It follows that $| \sum_{R_i \in W} Vol((R_i \times \mathbb{R}) \cap E) - Vol(E)| < \epsilon$.

We know $\int_D g - f = \sum_{R_i \in W} \int_{R_i} g - f + \sum_{R_i \in G \setminus W} \int g - f$, so $\int_D g - f = \sum_{R_i \in W} Vol((R_i \times \mathbb{R}) \cap$

$E) + \sum_{R_i \in G\setminus W} \int g - f$. Hence, $0 \le \sum_{R_i \in G\setminus W} \int g - f \le \sum_{R_i \in G\setminus W} M_i(g-f)|R_i| \le \sum_{R_i \in G\setminus W} (M_i(g) - m_i(f))|R_i| < \sum_{R_i \in G\setminus W} |R_i| \frac{\epsilon}{|R|} \le \epsilon$. This means that $|\sum_{R_i \in W} Vol((R_i \times \mathbb{R}) \cap E) - \int_E g - f| < \epsilon$, so $|Vol(E) - \int_E g - f| < 2\epsilon$. Since this is true for all $\epsilon > 0$ it follows that $\int_E g - f = Vol(E)$.

□

Note that with a simple re-labeling of axes (or by Theorems 12.60 and 12.57, interchanging coordinates) if D is a closed Jordan region the coordinates of which are listed in $(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_{n+1})$ then if f and g are continuous real valued functions on D then $E = \{(x_1, x_2, ..., x_j, ..., x_{n+1}) | (x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_{n+1}) \in D$ and $f(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_{n+1}) < x_j < g(x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_{n+1})\}$ is a Jordan region.

Setting up a double integral:

While a graph is not required, strictly speaking, it is a good idea to graph each domain over which we want to set up an iterated integral. The formula, as already described in the theorems above, is that if R is a type one region so that $R = \{(x, y) \in \mathbb{R}^2 | a \le x \le b$ and $g_1(x) \le y \le g_2(x)\}$ where g_1 and g_2 are continuous, then $\int_R f(x, y)dA = \int_a^b \int_{g_1(x)}^{g_2(x)} f(x, y)dydx$.

In practice, we usually think about the process of setting up an iterated integral as follows. First, we decide on a variable to be the "outer" variable, which is variable corresponding to the bounds in the leftmost integral sign. We then list that variable furthest to the right on the right side (so if x is the outer variable then the integral looks like $\int_a^b \int_{g_1(x)}^{g_2(x)} f(x, y)dydx$ for a type one region, whereas if y is the outer variable then the integral has form $\int_c^d \int_{h_1(y)}^{h_2(y)} f(x, y)dxdy$ for a type two region.

Often, we could use either variable order assuming we are willing to subdivide the region into smaller regions (possibly writing the integral as a sum of integrals). The choice of which variable to use as the outer variable in these cases generally comes down to convenience and familiarity, but in some cases choosing one order of integration makes a huge difference in the difficulty of the problem. Sometimes finding an antiderivative in one order is simple but very difficult in the other. Sometimes we can express an integral as a single integral in one order but must break the region into many separate integrals if we write the integral in the other order.

Once we have chosen an outer variable, the other variable is the (first) "inner" variable (though "outer" and "inner" variables are not a formally accepted terminology throughout texts in general). The bounds for the outer variable are always numerical. They range over all the values that variable can take on in points of R. In other words, the integral is taken over the projection of R onto that variable's coordinates. For instance, if x is the outer variable for a connected region R then you would have x start at the minimum a of all x values so that $(x, y) \in \mathbb{R}$, and then the upper bound of the integral would be b, which would be the maximum of all x-values so that $(x, y) \in \mathbb{R}$.

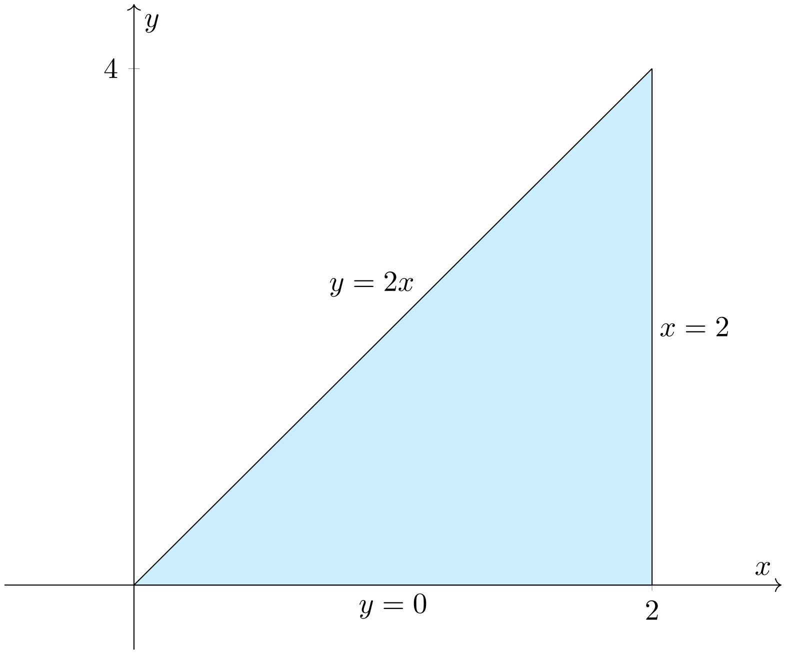
The bounds for the inner variable are not normally numbers. They are the functions $g_1(x)$ and $g_2(x)$. For any given x-value that means the bounds are numbers ($g_1(x)$ and $g_2(x)$ are numbers) but these numbers vary with x (because outside of the values between $g_1(x)$ and $g_2(x)$, the function f is zero). Thus, you think of the bounds of the outer variable as from the minimum to maximum value (numbers) for that variable, and think of the bounds for the inner variable (say it is x for this description) as having bounds $y = g_1(x)$ to $y = g_2(x)$ for each x in the outer variable range.

Here is an example:

Example 12.1. *Let R be the region bounded by the triangle with vertices $(0, 0)$, $(2, 0)$ and $(2, 4)$ in the plane. Find $\int_R xy^2 dA$, the volume under $f(x, y) = xy^2$ over the region R.*

Solution. We will choose x for the outer variable and y for the inner variable. We can see that the region R is bounded by the lines $y = 0$, $y = 2x$ and $x = 2$. This is easier to see from the graph below. Having determined this, the smallest value of x in the region is 0 and the largest value is 2. Over this interval $0 \leq x \leq 2$, the y values are between $y = 0$ and $y = 2x$ for each value of x. Hence, $\int_R xy^2 dA = \int_0^2 \int_0^{2x} xy^2 dydx = \int_0^2 \frac{xy^3}{3}\Big|_0^{2x} dx = \int_0^2 \frac{8x^4}{3} - 0dx = \frac{8x^5}{15}\Big|_0^2 = \frac{256}{15}$.

Graph of region R



Notice that we could have labeled the $y = 2x$ line as $x = \frac{y}{2}$ over $0 \leq y \leq 4$, the interval of values from the smallest to the largest values of y over the region. Note that the lower bound for the inner variable is $\frac{y}{2}$ and the upper bound is 2 because the integral's bounds are from the smallest value of of the variable within the region to the largest one (and $x = \frac{y}{2}$ is a smaller value of x than $x = 2$ over $y \in [0, 2]$. Hence, had we chosen y as the outer variable we could have set up the integral (and gotten the same answer, as guaranteed by Fubini's theorem) as follows:

$$\int_R xy^2 dA = \int_0^4 \int_{\frac{y}{2}}^2 xy^2 dxdy = \int_0^4 \frac{x^2y^2}{2}\bigg|_{\frac{y}{2}}^2 dy = \int_0^4 2y^2 - \frac{y^4}{8} dy = \frac{2y^3}{3} - \frac{y^5}{40}\bigg|_0^4 = \frac{128}{3} - \frac{1024}{40} = \frac{128}{3} - \frac{128}{5} = \frac{256}{15}.$$

□

It is frequently asked whether there is an algorithm for setting up an iterated integral that does not rely on knowing what the graph looks like. Such algorithms tend to only work for certain classes of cases and tend to be more difficult than looking at a graph. Part of the obstacle is trying to decide which points are in the region bounded by a set of curves. Technically, it does seem to be possible to describe an algorithm for doing this, but we are not aware of any algorithms that are not excessively complicated.

In some simpler cases, however one can see aspects of the region without a graph. If you can identify that a portion R of a region bounded by curves is the set of points whose x-values are between two values and whose y-values are between two functions then you can set up the bounds as we did above.

It is generally sensible to find points of intersection points of curves in a collection of curves that bound R and then try to express the region as the points between values of intersection points in one variable with the other variable a function between the lowest and highest values of that variable over the values of the outer variable.

Usually, graphing is actually not necessary to see the lowest and highest values of a variable in a region and then use the boundary curves to express the values in the region as between two functions of the outer variable. However, graphing is a good idea to help you check your work, and is a very helpful tool for seeing what the bound should be when it is not obvious from the boundary functions listed.

A useful observation that is so immediate that it probably doesn't deserve to be called a theorem (but which we will present as a theorem anyway so we can refer to it more easily) is that if an integrand can be written as a product of functions of single variables and the bounds of the iterated integral are all numerical, then the integral is the same as the product of the corresponding individual integrals in the corresponding variables.

Theorem 12.54. *Let $f_i(x_i)$ be continuous on $[a_i, b_i]$ for all $1 \leq i \leq n$ for some $n \in \mathbb{R}^n$, and let $R = \prod_{i=1}^{n}[a_i, b_i]$. Then $\int_R f_1(x_1)f_2(x_2)...f_n(x_n) =$*

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_n}^{b_n} f_1(x_1)f_2(x_2)...f_n(x_n)dx_n...dx_2dx_1 =$$

$$(\int_{a_1}^{b_1} f_1(x_1)dx_1)(\int_{a_2}^{b_2} f_2(x_2)dx_2)...(\int_{a_n}^{b_n} f_n(x_n)dx_n).$$

Proof. Since $f_i(x_i)$ for $1 \leq i \leq n-1$ are all constant when integrating with respect to x_n, we can move these functions outside of the integral sign for the integral with respect to x_n which gives us $\int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_n}^{b_n} f_1(x_1)f_2(x_2)...f_n(x_n)dx_n...dx_2dx_1$

$= \int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_{n-1}}^{b_{n-1}} f_1(x_1)f_2(x_2)...f_{n-1}(x_{n-1})(\int_{a_n}^{b_n} f_n(x_n)dx_n)dx_{n-1}...dx_2dx_1$, and then since $f_n(x_n)$ is constant relative to integration with respect to the other variables, we can take $\int_{a_n}^{b_n} f_n(x_n)dx_n$ out of the integral, giving us

$(\int_{a_n}^{b_n} f_n(x_n)dx_n)\int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_{n-1}}^{b_{n-1}} f_1(x_1)f_2(x_2)...f_{n-1}(x_{n-1})dx_{n-1}...dx_2dx_1$. We continue, taking out $\int_{a_{n-1}}^{b_{n-1}} f_{n-1}(x_{n-1})dx_{n-1}$ out of the integral and so on, ending with

$\int_R f_1(x_1)f_2(x_2)...f_n(x_n) = (\int_{a_1}^{b_1} f_1(x_1)dx_1)(\int_{a_2}^{b_2} f_2(x_2)dx_2)...(\int_{a_n}^{b_n} f_n(x_n)dx_n)$.

□

Here is an example of how this theorem can make iterated integrals simpler:

Example 12.2. *Find* $\int_0^1\int_0^2 x^2e^y dydx$.

Solution. Since the bounds are numerical and we can represent the integrand as a function of x (namely x^2) and a function of y (namely e^y), we can write $\int_0^1\int_0^2 x^2e^y dydx = (\int_0^1 x^2dx)(\int_0^2 e^ydy) = (\frac{x^3}{3}\Big|_0^1)(e^y\Big|_0^2) = (\frac{1}{3})(e^2-1) = \frac{e^2-1}{3}$.

□

The gains from using this trick are largely notational (we don't have to write as much on each step). This trick, while frequently used, only saves a little bit of time. We should be careful when using it that we do not attempt to use it at the wrong time. If the bounds are not numerical then it doesn't make sense, and if we can't write the integrand as a product of factors which are each single variable products then this doesn't make sense either.

Changing the order of integration:

Changing the order of integration is a method which is frequently used for simplifying an iterated integral. Sometimes an iterated integral does not have a nice antiderivative that can be found in the order in which the integral is written, but by expressing the bounds for the same region in another order the iterated integral becomes more easily integrated because the antiderivatives are much simpler. This is mainly helpful if the region is of type one and two (in general, it is more likely to be a useful method for iterated integrals if the iterated integrable is fully projectable with respect to multiple arrangements of the variables). Even if the region is not both type one and two, however, it can be useful if the region can be broken into smaller regions which are type one and two.

There doesn't really seem to be a nice algorithm which is universally applicable for reversing the order of integration in all cases. We recommend graphing the region expressed by an integral's bounds, and then looking at how to express the region as an iterated integral in the other variable order. However, assuming we have a type one and two region we can (sort of) describe an algorithm for reversing the order of integration without the use of a graph. The advantage to graphing is that it makes it easier to see whether the region is type one and type two and what the higher and lower function is over each interval (or whether the region would be better broken into multiple separate regions with separate iterated integrals).

The procedure when a region is both type one and type two is that you invert the functions representing the inner integral bounds if possible, determine the smallest and largest values of the inner variable, make those the bounds for the new outer variable and use the inverse functions as boundary curves to help you create bounds for the new inner variable integral bounds.

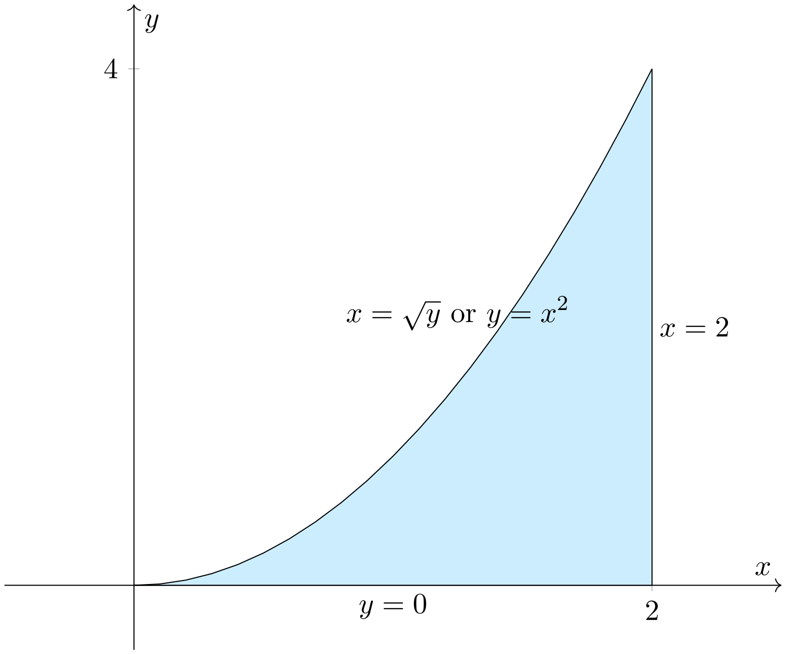
Here is an example:

Example 12.3. *Integrate* $\int_0^4 \int_{\sqrt{y}}^2 e^{x^3} dxdy.$

Solution. taking this integral in the current order won't work well. We don't have a nice antiderivative of e^{x^3} if we are integrating with respect to x. So, we will try switching the order of integration to see if that helps us evaluate this integral.

We will refer to R as the region described by the bounds of this integral, which is the region where $0 \leq y \leq 4$, and $\sqrt{y} \leq x \leq 2$ for each value of x. So, it is the region bounded on the left by $x = \sqrt{y}$ which is also the curve $y = x^2$ and bounded on the right by $x = 2$. This is the graph.

Graph of region R



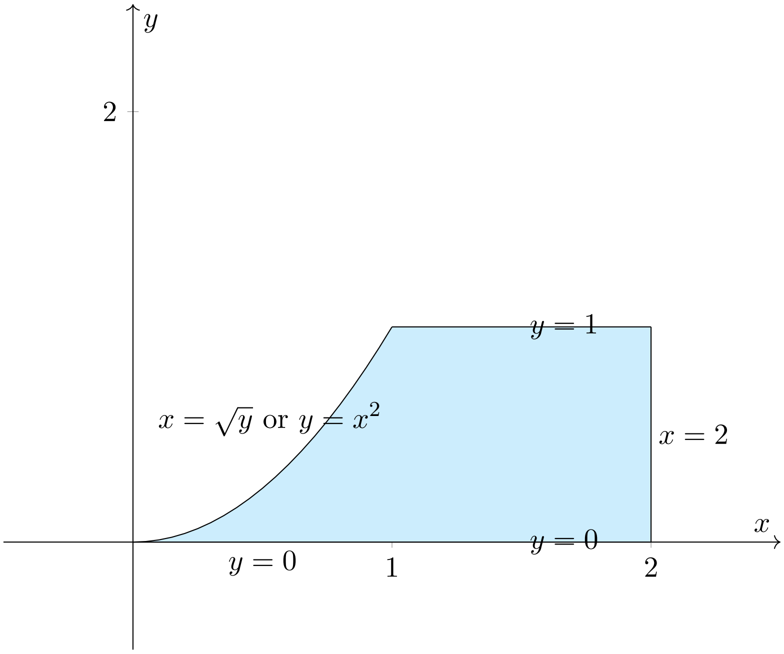
To set up the bounds for the integral in the other order, we must use x as the outer variable. We look for the smallest value of x in the region, which is 0, and the largest, which is 2, to get the bounds of the outer integral. Over $0 \leq x \leq 2$ we see that the curves bounding the region expressed as functions of x are $y = 0$ and $y = x^2$.

Hence, the integral becomes $\int_0^2 \int_0^{x^2} e^{x^3} dydx$. (Note that the integrand does not change when the order of integration is switched, only the bounds). This is $\int_0^2 ye^{x^3}\Big|_0^{x^2} = \int_0^2 x^2e^{x^3}dx$. Setting $u = x^3$ we have $du = 3x^2dx$, so the integral is $\frac{1}{3}\int_0^8 e^u du = \frac{e^u}{3}\Big|_0^8 = \frac{e^8-1}{3}$.

□

In the preceding example, suppose we didn't want to graph anything. Well, we could have said that the inverse of $x = \sqrt{y}$ is $y = x^2$, and we could have said that over $0 \leq y \leq 4$ the smallest and largest values of $x = \sqrt{y}$ were 0 and 2, which would be mean that we would have 0 and 2 as the bounds for the outer x-variable integral, and then we would go from $y = 0$ (since that is the smallest value for y) to $y = x^2$ for the inner variable. We have to be careful when using this process, though, because reversing the order of integration in general can't just be done by inverting functions, finding maxima and minima and then switching the integrals. Suppose, for instance, that the initial integral bounds had been $\int_0^1 \int_{\sqrt{y}}^2 f dxdy$ for a function f. Then the inverse of $x = \sqrt{y}$ would be $y = x^2$ as before, but because the function $x = \sqrt{y}$ does not reach $x = 2$ on $0 \leq y \leq 1$ we would have to express the integral with two integrals if we switch the variables. This would be the graph:

Graph of region R



So, in reversed order the integral would be $\int_0^1 \int_0^{x^2} f dydx + \int_1^2 \int_0^1 f dydx$. Depending on the integrand, this switching of variables to exchange one integral for two might be a mistake (this might make the problem harder).

A common error for readers who are not paying attention to the descriptions of what the bounds mean is to simply switch the integrals and keep the original variable representations of the bounds. This does not work. Consider the integral $\int_0^1 \int_x^1 e^y dydx$. If we just switched the integrals we would have $\int_x^1 \int_0^1 e^y dxdy$. The outer integral doesn't even have numerical bounds so the output of the calculation would have a variable in the answer (which is nonsense). The only time you can just switch the integral order and keep the same bounds as described is if the bounds are all numerical (meaning that you are integrating over a rectangle). It is correct that $\int_a^b \int_c^d f(x,y)dydx = \int_c^d \int_a^b f(x,y)dxdy$, which is probably part of the reason students of calculus sometimes think they can make the same sort of switch when there are non-numerical bounds.

We conclude this section by addressing how to find the volume between surfaces.

Finding the volume between two surfaces:

We proved in Theorem 12.53 that if $g(x,y) \leq f(x,y)$ over a Jordan region R then the volume of the region E between the graphs of $z = g(x,y)$ and $z = f(x,y)$ over the region R is $\int_R f(x,y) - g(x,y)dA$.

In some cases f and g might cross through each other, of course, in which case we would have to break R into sub-regions where one function is less than the other. Thus, in general, the volume between f and g over R is $V = \int_R |f(x,y) - g(x,y)|dA$.

Example 12.4. *Find the volume between the functions $g(x,y) = x + 2y$ and $f(x,y) = 9 - x^2 - y^2$ over the region R, where R is the region between the functions $y = 0$ and $y = x^2$ over $0 \leq x \leq 1$.*

Solution. This is $\int_0^1 \int_0^{x^2} 9 - x^2 - y^2 - x - 2ydydx$ since $f(x,y) > g(x,y)$ on all of R. This becomes $\int_0^1 9y - x^2y - \frac{y^3}{3} - xy - y^2\Big|_0^{x^2} dx = \int_0^1 9x^2 - x^4 - \frac{x^6}{3} - x^3 - x^4 dx = 3x^3 - \frac{x^4}{4} - \frac{2x^5}{5} - \frac{x^7}{21}\Big|_0^1 = \frac{967}{420}$.

□

Finding the average value of a function over a region:

Definition 95

The *average value* of an integrable function f over a Jordan region $E \subset \mathbb{R}^n$ is $\frac{1}{Vol(E)} \int_E f$.

Finding an average value is fairly straightforward (we just plug into the formula above). The average value is the height at which, if we took a constant function, the integral of the constant function would be the same as the integral of the original function. For a two variable function this is easy to visualize. If we pretend that ice does not change in volume when it melts and becomes water (which isn't true) then we could think of the volume under a surface over the region R in the xy-plane as being filled with ice. If the ice were then to melt but remain contained above the region R then the water level would be the average value of the function.

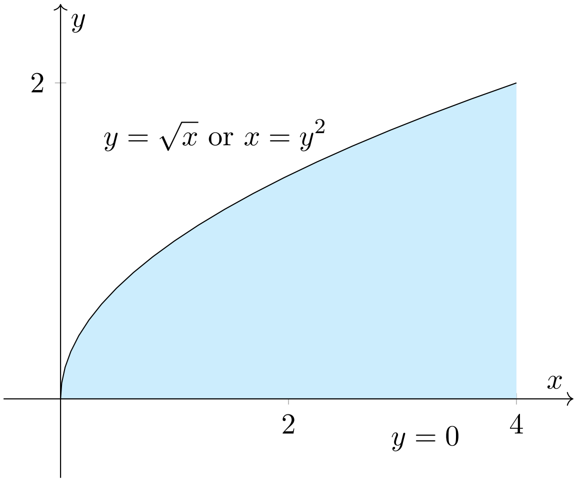
Example 12.5. *Find the average value of the function $f(x, y) = xy^2$ over the region R bounded by $y = \sqrt{x}$, $y = 0$ and $x = 4$.*

Solution. It would eliminate fractional powers to use y as the outer variable in this problem, which might be nice. So, the smallest value of y would be zero, and the largest would be two (the square root of four). Then $y = \sqrt{x}$ could be changed to $x = y^2$, so that $y^2 \leq x \leq 4$. Thus, the area of the region R would be $\int_0^2 4 - y^2 dy = 4y - \frac{y^3}{3}\Big|_0^2 = \frac{8}{3}$.

The average value would then be $\frac{3}{8}\int_0^2 \int_{y^2}^4 xy^2 dxdy = \frac{3}{8}\int_0^2 \frac{x^2y^2}{2}\Big|_{y^2}^4 = \frac{3}{8}\int_0^2 8y^2 - \frac{y^6}{2}dy = \frac{3}{8}(\frac{8y^3}{3} - \frac{y^7}{14})\Big|_0^2 = \frac{32}{7}$.

□

Graph of region R



Finally, though it is not a standard term, it is often helpful to describe functions satisfying the conditions in Exercise 12.1, so we make the following definition.

Definition 96

Let $f : E \to \mathbb{R}$ be integrable on a Jordan region E in $\mathbb{R}^n$ so that for some $i \in \{1, 2, 3, ..., n\}$, if $(x_1, x_2, ..., x_i, ..., x_n) \in E$ then $(x_1, x_2, ..., -x_i, ..., x_n) \in E$, and $f(x_1, x_2, ..., x_i, ..., x_n) = -f(x_1, x_2, ..., -x_i, ..., x_n)$. Then we say that f is *odd with* *respect to x_i* over region R which is *symmetric with respect to x_i.*

As shown in Exercise 12.1, the integral of a function f which is odd in a variable x_i over a region which is symmetric with respect to x_i is always zero.

Three dimensional integrals are of particular interest to us in applications, largely since our universe seems to be three dimensional, at least in terms of the dimensions associated with space. Most of the development is already done for us by the theorems in the earlier sections (since they were proven for an arbitrary dimension).

Cavalier's principle is an example of a notion established by triple integrals.

Theorem 12.55. *Cavalieri's Principle. Let E_1, E_2 be solids which are type 3 regions and are between two parallel planes $z = h$, $z = k$. If the areas of the cross sections at each height are the same then the volumes of the two solids are equal.*

Proof. Since, by Fubini's Theorem, we can write this volume as $\lim_{n\to\infty} \sum_{i=1}^{n} \Delta z \sum_{i=1}^{n} \int\int_E f(x, y, z_i^*) dA$, where $\int\int_E f(x, y, z_i^*) dA$ is the area of the cross section at height z_i^*, which is the same for both solids, the result follows. □

As verified in earlier sections, iterated integrals with three integrations can be used to evaluate integrals over fully projectable (type 3) regions in $\mathbb{R}^3$. If we integrate the number 1 over a solid E then the integral is the volume of E, though typically it makes more sense to do a double integral to determine a volume. However, many applications are based on integrating other functions over a solid. Any time you have something whose quantity per unit volume is known within a solid and you want to find the total amount of that thing over the solid you take a triple integral. For instance, if you know the mass per unit volume at each point within a solid then integrating that density function would give you the total mass of the solid. If you know the probability that an outcome will occur per unit volume within a solid then the integral over the solid will give the probability that the outcome will occur somewhere within the solid. If you know the total amount of heat energy per unit volume then the integral over the solid would give you the total heat energy within the solid. If you know the total charge per unit volume (or the force acting on a particle from this charge) then the integral would give the total charge (or total force on the particle) from the entire solid region. If you know the number of molecules of a chemical per cubic unit of volume within a solid region (perhaps a large portion of gas like the atmosphere over a city) then integrating would give you the total number of molecules of the chemical within the region.

.

Setting up triple integrals:

Iterated integrals with three integral signs follow a similar pattern to double integrals, but the graphing and interpreting the graph tends to be more difficult. It is usually good to begin by looking at the graph of the solid E to be integrated over and looking for a direction onto which the solid, if projected onto the coordinate plane perpendicular to that direction, projects to a type one or two region D on a coordinate plane. So, if we call that direction z (it could be x or y, of course) then there are continuous functions $h_1(x,y) \leq h_2(x,y)$ defined on D so that E is the set of all points (x,y,z) so that $(x,y) \in D$ and $h_1(x,y) \leq z \leq h_2(x,y)$. We then look at D and set up bounds for the region as we did for a double integral over D, selecting an outer variable to have numerical bounds and then letting the other variable's integral bounds be functions of the first variable. Note that at the end of this process the outer variable is numerical the second integral bounds are functions of the outer variable and the third (inner) integral bounds are functions of the preceding two variables. It may help some people to thing of the innermost integral as the integrand function for a double integral, so $\int\int\int_E f dV = \int\int_D (\int_{h_1(x,y)}^{h_2(x,y)} f) dA$.

The process of integrating is the same idea as a double integral. Specifically, we treat all other variables to be constant except the variable we are integrating with respect to in each integral, plug in the bounds and then move to the next integral until all integrals are completed. This is illustrated in the following example:

Example 12.6. *Find* $\int_0^1 \int_0^{x^2} \int_0^{xy} zx dz dy dx$.

Solution. We integrate with respect to z then y and then x as follows: $\int_0^1 \int_0^{x^2} \int_0^{xy} zx dz dy dx = \int_0^1 \int_0^{x^2} \frac{z^2 x}{2}\Big|_0^{xy} dy dx = \int_0^1 \int_0^{x^2} \frac{x^3 y^2}{2} dy dx = \int_0^1 \frac{x^3 y^3}{6}\Big|_0^{x^2} dx = \int_0^1 \frac{x^9}{6} dx = \frac{x^{10}}{60}\Big|_0^1 = \frac{1}{60}$. □

It is frequently possible to switch the order of the variables by changing the outer variable or the projected region D. The integrand is unaffected by this process of setting up the bounds. Here is an example where the bounds can be set up in six different ways because the solid is fully projectable in any variable order. We will set up three of these integrals.

Example 12.7. *Let* $I = \int_E f$*, where* E *is the region bounded by the surfaces* $z = 0$*,* $y = 9 - x^2$ *and* $z = y$*. Set up bounds to evaluate this integral in three different ways.*

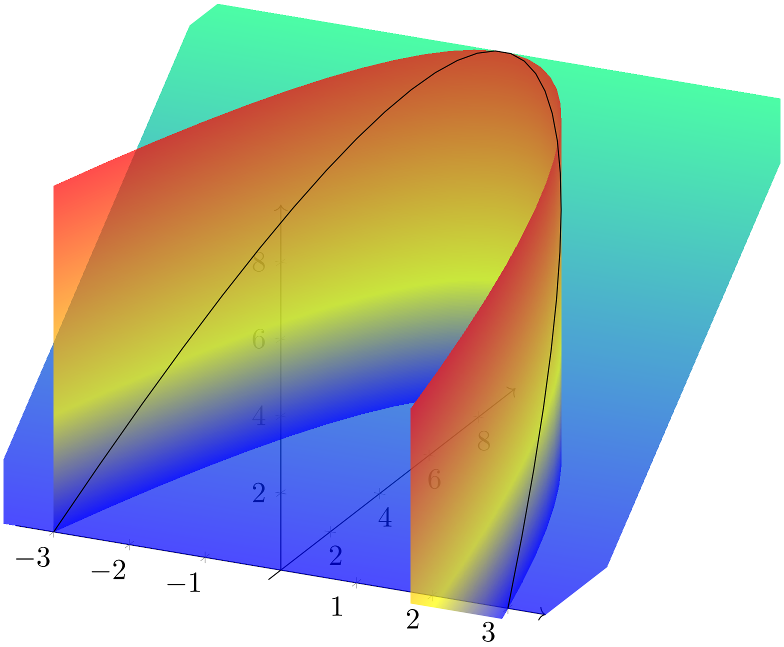
Solution. It is helpful to graph these solids before setting up the triple integral, so we include a graph below. We can see that if $z = 0$ and $z = y$ then $y = 0$ at the intersection of these surfaces. We can also see that if $y = 0$ and $y = 9 - x^2$ then the intersection of these three surfaces occurs when $x = \pm 3$.

We will start by viewing the projection D_{xy} of this solid onto the xy-plane as the region between the line $y = 0$ and the parabola $y = 9 - x^2$. Then the outer variable could be x, where $-3 \leq x \leq 3$. The next variable could be y, where $0 \leq y \leq 9 - x^2$, which gives the region D_{xy}. We could then have $0 \leq z \leq y$ over D. This gives us $\int_{-3}^{3} \int_0^{9-x^2} \int_0^y f dz dy dx$.

One way to set the integral up with different bounds is just to use the same projection D_{xy} and reverse the order of integration on that region. This is two dimensional region, so it is a process we are already familiar with, and for each $0 \leq y \leq 9$ we would have $-\sqrt{9-y} \leq x \leq \sqrt{9-y}$, so the integral would be $\int_0^9 \int_{-\sqrt{9-y}}^{\sqrt{9-y}} \int_0^y f dz dy dx$.

For our third choice of bounds we will project E onto the yx-plane to a region D_{yz}, which is a region bounded by the triangle with sides $z = y$, $z = 0$ and $y = 0$. The largest value of y for this triangle is $y = 9$, and for every value of y we would have $0 \leq z \leq y$. The function x is restricted only by y in the second surface, so $-\sqrt{9-x^2} \leq x \leq \sqrt{9-x^2}$. We can thus express the integral as $\int_0^9 \int_0^y \int_{-\sqrt{9-y}}^{\sqrt{9-y}} f\,dxdzdy$.

Graph of E



□

Transformation of Variables

It is frequently useful to take a particular coordinate system and change it to another, such as changing from rectangular to polar coordinates. We would like to be able to make such substitutions in a manner which is easy to manage. The following theorem is somewhat tricky to prove, and the arguments we give may seem a bit cumbersome. There are several slightly different forms of this result. For now, we are just stating the theorem.

Theorem. Transformation of Variables. *Let $\phi : U \to \mathbb{R}^n$ be a one to one C^1 function, where U is open in $\mathbb{R}^n$, $\det D\phi(\boldsymbol{x}) \neq 0$ on U and E is a Jordan region whose closure is contained in U. Then* $\int_{\phi(E)} f = \int_E f \circ \phi |\det D\phi|$.

This theorem is helpful because it allows us to justify many conversions of variables formally to coordinate systems like polar coordinates, spherical coordinates and cylindrical coordinates (the latter two are discussed in the next section). This is much more general, however, and allows us to transform more complicated regions into simpler ones if we can come up with the right transformations to make this work.

We first discuss why this theorem is reasonable. We will discuss this as a sequence of observations.

First, observe that due to the Inverse Function Theorem, a point $\mathbf{x}$ in the interior of E will be contained in an open set contained in E whose image contains an open ball which contains $\phi(\mathbf{x})$. Since ϕ is one to one it must follow that the image of any open ball on the boundary of E contains a point in $\phi(E)$ and a point not in $\phi(E)$, which means that the boundary of $\phi(E)$ is $\phi(\partial(E))$.

Second, focusing on the two dimensional case for now, let $\phi(u,v) = (x(u,v), y(u,v))$ in $\mathbb{R}^2$. If we were to take the edges of a rectangle $R_{ij} = [u_i, u_i + \Delta u] \times [v_j, v_j + \Delta v] \subset E$ with vector sides $< \Delta u, 0 >$ and $< 0, \Delta v >$ then if the derivative of the transformation ϕ were a *constant* matrix $D = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$ on the rectangle, we would have that the function $\phi(u_i + h_1, v_j + h_2) - \phi(u_i, v_j) = (\frac{\partial x}{\partial u}h_1 + \frac{\partial x}{\partial v}h_2, \frac{\partial y}{\partial u}h_1 + \frac{\partial y}{\partial v}h_2)$ by Taylor's Theorem (or the Mean Value Theorem for Real Valued Functions for that matter) on each component function. If we only look at points on the rectangle then $0 \leq h_1 \leq \Delta u$ and $0 \leq h_2 \leq \Delta v$. This means that the image of the rectangle R_{ij} (which is, by definition, $\{(u_i + t_1\Delta u, v_j + t_2\Delta v)|0 \leq t_1 \leq 1 \text{ and } 0 \leq t_2 \leq 1\}$) under ϕ would be the parallelogram $P_{ij} = \{(x(u_i, v_j) + \frac{\partial x}{\partial u}\Delta u t_1 + \frac{\partial x}{\partial v}\Delta v t_2, y(u_i, v_j) + \frac{\partial y}{\partial u}\Delta u t_1 + \frac{\partial y}{\partial v}\Delta v t_2)|0 \leq t_1 \leq 1 \text{ and } 0 \leq t_2 \leq 1\}$. In other words, the image of the rectangle would be a parallelogram. A similar argument would have shown the image of a three dimensional rectangle would have been a three dimensional parallelpiped. Note that the vector sides of the new parallelogram would be $(\frac{\partial x}{\partial u}\Delta u, \frac{\partial y}{\partial u}\Delta u)$ and $(\frac{\partial x}{\partial v}\Delta v, \frac{\partial y}{\partial v}\Delta v)$.

Third, we notice that if the derivative is not constant, but is very close to constant (each partial is between two very close values) on the rectangle R_{ij} then $\phi(R_{ij})$ is contained between two parallelograms $P_{ij} \subseteq R_{ij} \subseteq Q_{ij}$ that are very close to one another. In fact, they are so close that the areas satisfy the condition that the ratio $\frac{Vol(P_{ij}) - Vol(Q_{ij})}{Vol(R_{ij})}$ can be made as small as we wish, as long as the derivatives are bounded between sufficiently close values. In other words, if we were to take the derivative at any point $\mathbf{p}_{ij} \in R_{ij}$ and use this to find a parallelogram M_{ij} that would have been the image of R_{ij} had the matrix $D\phi(\mathbf{p}_{ij})$ been equal to $D\phi(\mathbf{x})$ for all $\mathbf{x} \in R$ then we can make $\frac{Vol(M_{ij})}{Vol(\phi(R_{ij}))}$ as close to one as we wish. We can also get a similar same result for parallelpipeds. These are both based on the idea of establishing that if derivatives are near constant then the images of line segments are near the starting points plus the displacement vectors acted on by the linear transformations of those segments (the images had the derivative been the constant the derivative is near), in that the lengths of displacements between vectors are similar (their ratio is near one) and the angles between the displacement vectors are small. This can let us show that the image fits within the desired parallelograms.

Fourth, recall that we discussed that the area of a parallelogram in $\mathbb{R}^2$ with vector edges $\mathbf{u}, \mathbf{v}$ is $|\det(\mathbf{u}, \mathbf{v})|$ and that the volume of a parallelpiped with vector edges $\mathbf{u}, \mathbf{v}, \mathbf{w}$ is

$|\det(\mathbf{u}, \mathbf{v}, \mathbf{w})|$, which was demonstrated in Theorem 10.4 under some minimal assumptions of geometry. This means that in the previous step the volume $Vol(M_{ij}) = \Delta x \Delta y |\det D\phi(\mathbf{p}_{ij})|$. We have a more careful form of this observation for n dimensions below (not relying on geometric concepts of volume that we haven't developed with proper rigor).

Fifth, we can show that if the base rectangle were subdivided into a union of non-overlapping Jordan regions (instead of a grid consisting of subrectangles) then multiplying the volumes of the Jordan regions times the function values would give a sum that can be made as close as we wish to the integral if the diameters of those Jordan regions are sufficiently small.

Finally, we can take a grid so that the rectangles on the grid which intersect the region E and their corresponding images under ϕ have very small diameter. Since the partial derivatives are all continuous on the closed bounded rectangle they are uniformly continuous, so we can subdivide the grid into small enough rectangles so that the derivatives are close enough to constant on each subrectangle that the ratios of the $\dfrac{Vol(M_{ij})}{Vol(\phi(R_{ij}))}$ are close to one on every rectangle, meaning that $\Delta x \Delta y = k_{ij} \Delta x \Delta y |\det D\phi(\mathbf{p}_{ij})|$ where k_{ij} is close to one for all i, j. Thus, if we let S be the set of rectangles in the interior of E in this subdivision with fine mesh then $\displaystyle\sum_{R_{ij} \in S} f(\phi(\mathbf{p}_{ij}))|R_{ij}||\det D\phi(\mathbf{p}_{ij})|$ is very close to $\displaystyle\sum_{\{\phi(R_{ij})|R_{ij} \in S\}} f(\mathbf{q}_{ij}) Vol(\phi(R_{ij}))$ where the $\mathbf{p}_{ij}$ points are chosen as the preimages of the $\mathbf{q}_{ij}$ (so $f(\mathbf{q}_{ij})$ means the same thing as $f(\phi(\mathbf{p}_{ij}))$. Each of these sums can be made arbitrarily close to the two integrals in the theorem listed above, which means that the integrals are equal.

This argument makes the result above make sense, but it is just an intuitive description and ideas like making volumes close over sums when derivatives are close takes some effort to formalize.

We can use the change of variables formula to get formulas for integrals with polar coordinates. The maps $x(r, \theta) = r\cos(\theta)$ and $y(r, \theta) = r\sin(\theta)$ are C^1 maps on $[0, R] \times [0, 2\pi]$ and these maps are one to one except on a set of volume zero. Hence, using the change of variables formula, if R is a region in the xy-plane which is the image of D in the $r\theta$ plane, we have $\displaystyle\int\int_R f(x, y) dA = \int\int_D f(r\cos\theta, r\sin\theta)|\det J| dA$, where $J = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix}$, so $\det J = r\cos^2(\theta) + r\sin^2(\theta) = r$. From this it is possible to more formally derive the formula for polar area in one variable. If we want to find the area enclosed by $r = r(\theta)$ and the origin over $\theta_1 \leq \theta \leq \theta_2$, we can call this region R and then the area A of R is found by integrating 1 over R, which means that $A = \displaystyle\int\int_R 1 dA = \int_{\theta_1}^{\theta_2} \int_0^{r(\theta)} 1(r) dr d\theta$ $= \displaystyle\int_{\theta_1}^{\theta_2} \frac{r^2}{2}\Big|_0^{r(\theta)} d\theta = \frac{1}{2}\int_{\theta_1}^{\theta_2} (r(\theta))^2 d\theta$, which is the formula for polar area we used in chapter six. Since this development is more formal and less pictorial (with a clear definition for what area means rather than relying as heavily on properties of geometry that we have not derived), this is a more rigorous derivation of this formula.

Example 12.8. *Find* $\displaystyle\int\int_E x^2 dA$*, where* E *is the region bounded by the ellipse* $\dfrac{x^2}{4} + \dfrac{y^2}{9} = 1$.

Solution. If we set $x = 2u$ and $y = 3v$ then we would have $\frac{4u^2}{4} + \frac{9v^2}{9} = u^2 + v^2 = 1$. Thus, the transformation described takes the unit disk D to the elliptic disk E. The Jacobian of the transformation is $\det \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = 6$. Thus, the integral becomes $\int\int_D 6(4u^2)dA$. We can than change the integral to polar coordinates so that it becomes $24\int_0^{2\pi}\int_0^1 r^3\cos^2(\theta)drd\theta = 24\int_0^{2\pi}\cos^2(\theta)d\theta\int_0^1 r^3dr = 24(4)(\frac{1}{2})(\frac{\pi}{2})\frac{r^4}{4}\Big|_0^1 = 6\pi.$

□

Example 12.9. *Find $\int\int_E 4x^2 + yxdA$, where E is the parallelogram bounded by the lines $4x + y = 0$, $4x + y = 2$, $x - y = 1$ and $x - y = 4$.*

Solution. We will set $u = 4x + y$ (which ranges from 0 to 2 on the parallelogram) and $v = x - y$ (which ranges from 1 to 4 on the parallelogram). We solve for x and y, beginning with adding these equations to get $5x = u + v$, so $x = \frac{u+v}{5}$. Subtracting four times the second equation from the first would give $5y = u - 4v$, so $y = \frac{u-4v}{5}$. This lets us get the Jacobian $\det \begin{bmatrix} \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & -\frac{4}{5} \end{bmatrix} = -\frac{1}{5}$, which has absolute value $\frac{1}{5}$. Note that the integrand is $(4x + y)(x) = u(\frac{u+v}{5})$. Thus, the integral becomes $\frac{1}{25}\int_0^2\int_1^4 u^2 + uvdvdu = \frac{1}{25}\int_0^2 u^2v + \frac{uv^2}{2}\Big|_1^4 du = \frac{1}{25}\int_0^2 4u^2 + 8u - (u^2 + \frac{u}{2})du = \frac{1}{25}\int_0^2 3u^2 + \frac{15u}{2}du = \frac{1}{25}(u^3 + \frac{15u^2}{4})\Big|_0^2 = \frac{23}{25}.$

□

Example 12.10. *Find $\int\int\int_E 2xzdV$, where E is the parallelpiped bounded by the planes $x+y+z = 0$, $x+y+z = 2$, $x-y-z = 0$ and $x-y-z = 1$, $x-y+z = 0$ and $x-y+z = 4$.*

Solution. We will set $u = x + y + z$ (which ranges from 0 to 2 on the parallelpiped) and $v = x - y - z$ (which ranges from 0 to 1 on the parallelpiped) and $w = x - y + z = 0$ (which ranges from 0 to 4 on the parallelpiped). We then solve for the variables. Adding the first two equations gives $u + v = 2x$, so $x = \frac{u+v}{2}$. Subtracting the third equation from the second gives $v - w = -2z$, so $z = \frac{w-v}{2}$. Subtracting the third equation from the first gives $u - w = 2y$, so $y = \frac{u-w}{2}$. The Jacobian determinant is then $\det \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = -\frac{1}{8} - \frac{1}{8} = -\frac{1}{4}$, the absolute value of which is $\frac{1}{4}$. Thus,

$\int\int\int_E 2xzdV = \int_0^2\int_0^1\int_0^4 (\frac{1}{4})(2)(\frac{u+v}{2})(\frac{w-v}{2})dwdvdu = \frac{1}{8}\int_0^2\int_0^1\int_0^4 uw - uv + vw - v^2dwdvdu = \frac{1}{8}(\int_0^2 udu\int_0^1 1dv\int_0^4 wdw + \int_0^2 1du\int_0^1 vdv\int_0^4 wdw - \int_0^2 udu\int_0^1 vdv\int_0^4 1dw - \int_0^2 1du\int_0^1 v^2dv\int_0^4 1dw) = \frac{1}{8}[(\frac{2^2}{2})(1)(\frac{4^2}{2}) + (2)(\frac{1^2}{3})(\frac{4^2}{2}) - (\frac{2^2}{2})(\frac{1^2}{2})(4) - (2)(\frac{1^3}{3})(4)] = 2 + \frac{2}{3} - \frac{1}{2} - \frac{1}{3} = \frac{11}{6}.$

□

Theorem 12.56. *Let $\phi : U \to \mathbb{R}^n$ be C^1, where U is an open set in $\mathbb{R}^n$ containing a compact set K, and $\Delta_\phi \neq 0$ on U. Then:*

(a) There is an $M > 0$ so that $L(\boldsymbol{x}, \boldsymbol{y}) \subseteq K$ then $|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| \leq M|\boldsymbol{x} - \boldsymbol{y}|$.

(b) Let $Vol(E) = 0$ where $\overline{E} \subseteq U$. Then $Vol(\phi(E)) = 0$. Furthermore, there is a $\delta > 0$ and a constant $C > 0$ so that if a cube W has diameter less than δ and intersects $\overline{E}$ then $\phi(W)$ is contained in a cube of volume $CVol(W)$.

Proof. (a) By the Mean Value Theorem for Vector Valued Functions, if $L(\mathbf{x}, \mathbf{y}) \subset K \subset U$ then there is some $\mathbf{c} \in L(\mathbf{x}, \mathbf{y})$ so that $(\phi(\mathbf{x}) - \phi(\mathbf{y})) \cdot D\phi(\mathbf{c})(\mathbf{x} - \mathbf{y}) = |\phi(\mathbf{x}) - \phi(\mathbf{y})|^2$.

By the Cauchy Schwartz Inequality and Theorem 10.32, $|\phi(\mathbf{x}) - \phi(\mathbf{y})|^2 \leq |\phi(\mathbf{x}) - \phi(\mathbf{y})| \cdot |D\phi(\mathbf{c})||\mathbf{x} - \mathbf{y}|$, so $|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq \cdot |D\phi(\mathbf{c})||\mathbf{x} - \mathbf{y}|$. Also by Theorem 10.32, for each $\mathbf{x} \in U$ we know $|D\phi(\mathbf{x})| \leq \sqrt{n}\sum_{i=1}^{n}\sum_{j=1}^{n} |\frac{\partial \phi_i}{\partial x_j}(\mathbf{x})|$, which is continuous since ϕ is C^1. By the Extreme Value Theorem, it follows that $\sqrt{n}\sum_{i=1}^{n}\sum_{j=1}^{n} |\frac{\partial \phi_i}{\partial x_j}(\mathbf{x})|$ takes on a maximum value M on K and therefore $|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$.

(b) For each $\mathbf{x} \in \overline{E}$ we can find $\gamma_{\mathbf{x}} > 0$ so that $\overline{C_{\gamma_{\mathbf{x}}}(\mathbf{x})} \subset U$. Since $\overline{E}$ is compact and $\{C_{\gamma_{\mathbf{x}}}(\mathbf{x})\}_{\mathbf{x} \in \overline{E}}$ is an open cover of $\overline{E}$, we can find a finite subcover $F = \{C_{\gamma_{\mathbf{x}_i}}(\mathbf{x}_i)\}_{1 \leq i \leq k}$, and we note that $K = \bigcup_{i=1}^{k} \overline{C_{\gamma_{\mathbf{x}_i}}(\mathbf{x}_i)}$ is compact and $\overline{E} \subseteq K \subset U$. By part (a) we can find $M > 0$ so that $|\phi(\mathbf{x}) - \phi(\mathbf{y})| \leq M|\mathbf{x} - \mathbf{y}|$ if $L(\mathbf{x}, \mathbf{y}) \subseteq K$.

Let $\epsilon > 0$. By the Lebesgue Number Lemma we can find $\delta > 0$ so that if a set S intersects $\overline{E}$ and has diameter less or equal to δ then S is a subset of an element of F. By Theorem 12.27, we can find $0 < \eta < \frac{\delta}{\sqrt{n}}$ and a collection of cubes $Q = \{Q_i\}_{1 \leq i \leq t}$ which covers $\overline{E}$, each of which has side length η and therefore diameter $\eta\sqrt{n} < \delta$ by Theorem 12.1, so that $\sum_{i=1}^{t} |Q_i| < \frac{\epsilon}{4^n M^n n^{\frac{n}{2}}}$.

If W is any cube with $diam(W) < \delta$ of side length L so that W intersects $\overline{E}$, since $W \subseteq C_{\gamma_{\mathbf{x}_j}}(\mathbf{x}_j)$ for some $j \in \{1, 2, 3, ..., k\}$, we know that W is a convex subset of K, so for any $\mathbf{x}, \mathbf{y} \in W$ it follows that $|\phi(\mathbf{x}) - \phi(\mathbf{y})| < M\sqrt{n}L$. Hence, for any point $\mathbf{z} \in \phi(W)$, we know that $B_{2M\sqrt{n}L}(\phi(\mathbf{z}))$ contains $\phi(Q)$ and hence by Theorem 12.1, it follows that $\phi(Q) \subseteq C_{4M\sqrt{n}L}(\phi(\mathbf{z}))$. The volume of W is L^n and the volume of $C_{4M\sqrt{n}L}(\phi(\mathbf{z}))$ is $(4M\sqrt{n})^n L^n = CVol(W)$, where $C = (4M\sqrt{n})^n$.

For each $1 \leq i \leq t$, the $diam(Q_i) < \delta$. For each $1 \leq i \leq t$ we can choose a point $\mathbf{z}_i \in \phi(Q_i)$. It follows that $\phi(Q_i) \subseteq C_{4M\eta\sqrt{n}}(\phi(\mathbf{z}_i))$, so $\phi(E) \subset \bigcup_{i=1}^{t} C_{4M\eta\sqrt{n}}(\mathbf{z}_i)$.

We know that $|Q_i| = \eta^n$, and $|C_{4M\eta\sqrt{n}}(\phi(\mathbf{z}_i))| = 4^n M^n n^{\frac{n}{2}} \eta^n = 4^n M^n n^{\frac{n}{2}} |Q_i|$. Thus, $\sum_{i=1}^{t} |C_{4M\eta\sqrt{n}}(\mathbf{z}_i)| = 4^n M^n n^{\frac{n}{2}} \sum_{i=1}^{t} |Q_i| < 4^n M^n n^{\frac{n}{2}} \frac{\epsilon}{4^n M^n n^{\frac{n}{2}}} = \epsilon$. Hence, $Vol(\phi(E)) = 0$.

□

First we note some properties of nice maps which are one to one, continuously differentiable and have non-zero derivative determinants.

Theorem 12.57. *Let $\phi : U \to \mathbb{R}^n$ be one to one and C^1, with $\Delta_\phi \neq 0$ on U, where U is an open set in $\mathbb{R}^n$ containing $\overline{E}$ and $E \subseteq \mathbb{R}^n$. Then $\phi(\partial(E)) = \partial(\phi(E))$.*

Furthermore, $\phi(\overline{E}) = \overline{\phi(E)}$.

Proof. By Theorem 11.24, ϕ is a homeomorphism.

Let $\mathbf{p} \in \partial(E)$. Let U be an open set containing $\mathbf{p}$. Then U contains a point $\mathbf{x} \in E$ and a point $\mathbf{y} \notin E$, so $\phi(U)$ contains a point $\phi(\mathbf{x}) \in \phi(E)$ and a point $\phi(\mathbf{y}) \notin \phi(E)$. Since every open set V containing $\phi(\mathbf{p})$ contains the image of an open set $\phi^{-1}(V)$, it follows that every open set containing $\phi(\mathbf{p})$ contains a point of $\phi(E)$ and a point not in $\phi(E)$, so $\phi(\mathbf{p}) \in \partial(\phi(E))$.

Since ϕ^{-1} is also continuous on the open set $\phi(U)$, it follows from the same argument that if $\mathbf{q} \in \partial(\phi(E))$ then $\phi^{-1}(\mathbf{q}) \in \partial(\phi^{-1}(\phi(E))) = \partial(E)$. Hence, $\phi(\partial(E)) = \partial(\phi(E))$.

From this we see that $\phi(\overline{E}) = \phi(E) \cup \phi(\partial(E)) = \phi(E) \cup \partial(\phi(E)) = \overline{\phi(E)}$.

□

Theorem 12.58. *(a) Let $\phi : U \to \mathbb{R}^n$ be C^1, where U is an open set in $\mathbb{R}^n$ containing the closure of Jordan region E, and let $\Delta_\phi \neq 0$ on U. Then $\phi(E)$ is a Jordan region.*

(b) If ϕ is also one to one on U except on a set S of volume zero, and E and D are non-overlapping Jordan regions whose closures are contained in U, then $\phi(E)$ and $\phi(D)$ are non-overlapping Jordan regions.

Proof. (a) By the Inverse Function Theorem we know that $\phi(V)$ is open for every open $V \subset U$. In particular, $\phi(E^\circ)$ is an open subset of $\phi(E)$, and is therefore contained in the interior of $\phi(E)$. Since E is a Jordan region we also know that E is bounded, so $\overline{E}$ is compact. Thus, $\phi(\overline{E})$ is compact since the continuous image of a compact set is compact, and is closed by the Heine-Borel Theorem, and contains $\phi(E)$. Since $\overline{\phi(E)}$ is the intersection of all closed sets containing $\phi(E)$ we know that $\overline{\phi(E)} \subseteq \phi(\overline{E})$. Thus, $\partial(\phi(E)) = \overline{\phi(E)} \setminus \phi(E)^\circ \subseteq \phi(\overline{E}) \setminus \phi(E^\circ) \subseteq \phi(\partial(E))$. Since E is a Jordan region, we know that $Vol(\partial(E)) = 0$, so by Theorem 12.56, it follows that $Vol(\phi(\partial(E))) = 0$ and thus $Vol(\partial(\phi(E))) = 0$, which means that E is a Jordan region.

(b) Since E and D are non-overlapping we know that $Vol(E \cap D) = 0$, which means that $Vol(\phi(E \cap D)) = 0$. Likewise, we know that $Vol(\phi(S)) = 0$. Let $\mathbf{y} \in (\phi(E) \cap \phi(D)) \setminus \phi(S)$. Then there is a unique point $\mathbf{x} \in U$ so that $\phi(\mathbf{x}) = \mathbf{y}$, which means that $\mathbf{x} \in E \cap D$. It follows that $(\phi(E) \cap \phi(D)) \setminus \phi(S) \subseteq \phi(E \cap D)$ and thus $(\phi(E) \cap \phi(D)) \subseteq (\phi(S) \cup \phi(E \cap D))$. Since the union of two sets of volume zero has volume zero, we know that $Vol(\phi(E) \cap \phi(D)) = 0$ and therefore $\phi(E)$ and $\phi(D)$ are non-overlapping.

□

Theorem 12.59. *(a) Let R be a rectangle in $\mathbb{R}^n$ and let $\epsilon > 0$. Then there is a collection Q of cubes whose interiors cover R so that $Vol(\bigcup Q \setminus R) < \epsilon$. Likewise, there is a collection W of non-overlapping cubes contained in R so that $Vol(R \setminus \bigcup W) < \epsilon$.*

(b) A set S in $\mathbb{R}^n$ has Lebesgue measure zero if and only if for every $\epsilon > 0$ there is a collection of cubes $\{C_i\}_{i\in\mathbb{N}}$ whose interiors cover S so that $\sum_{i=1}^{\infty} |C_i| < \epsilon$.

(c) Let $\phi : U \to \mathbb{R}^n$ be one to one, C^1 and have $\Delta)\phi \neq 0$ on U. Let E be a Jordan region so that $\overline{E} \subset U$ and $\lambda(E) = 0$. Then $\lambda(\phi(E)) = 0$.

Proof. (a) First, let K be a cube containing R. By Theorem 12.29 we know that there is a $\delta > 0$ so that if a grid G on K has mesh less than δ then for the function $f(\mathbf{x}) = 1$ on R we have $U(f, G) - L(f, G) = V(R, G) - v(R, G) < \epsilon$. Let $G = \{R_i\}_{1\leq i\leq k}$ be a grid on K consisting of cubes (obtained by dividing each edge factor of K into the same number of evenly spaced subdivisions) with $|G| < \delta$. Then we know that $Vol(R) - \epsilon < v(R, G) \leq Vol(R) \leq V(R, G) < Vol(R) + \epsilon$. By Theorem 12.7 we can replace each $R_i \in G$ be a cube Q_i so that $R_i \subset Q_i^\circ$ and $Vol(Q_i) - Vol(R_i) < \dfrac{\epsilon - (V(R, G) - Vol(R))}{k}$, so that $\sum_{i=1}^{k} |Q_i| < Vol(R) + \epsilon$ and $R \subset \bigcup_{i=1}^{k} Q_i^\circ$.

By Theorem 12.7.

(b) If there is a such a collection of cubes for each $\epsilon > 0$ then by definition we know that $\lambda(S) = 0$. Assume that $\lambda(S) = 0$ and let $\epsilon > 0$. Then we can pick a collection of rectangles $\{R_i\}_{i\in\mathbb{N}}$ covering S so that $\sum_{i=1}^{\infty} |R_i| < \frac{\epsilon}{2}$. By (a), for each i we can pick a collection $C_i = \{W_j\}_{1\leq j\leq n_i}$ of cubes whose interiors cover R_i so that $Vol(\bigcup C_i \setminus R_i) < \frac{\epsilon}{2^{i+1}}$. Hence, $\sum_{i=1}^{\infty} \sum_{j=1}^{n_i} |W_j| < \sum_{i=1}^{\infty} \sum_{i=1}^{\infty} |R_i| + \frac{\epsilon}{2^{i+1}} < \epsilon$.

(c) By Theorem 12.56 we know that there is a $\delta > 0$ and a $C > 0$ so that if W is a cube of diameter less than δ and $W \cap \overline{E} \neq \emptyset$ then $\phi(W)$ is a contained in a cube of volume no larger than $CVol(W)$.

Let $\lambda(E) = 0$. Then we can find a collection of cubes $\{Q_i\}_{i\in\mathbb{N}}$ which covers E so that the diameter of each Q_i is less than δ and $\sum_{i=1}^{\infty} |Q_i| < \frac{\epsilon}{C}$. For each Q_i we choose a cube R_i containing $\phi(Q_i)$ so that $Vol(R_i) \leq CVol(W)$. Then $\{R_i\}_{i\in\mathbb{N}}$ covers $\phi(E)$ and $\sum_{i=1}^{\infty} |R_i| < C\frac{\epsilon}{C} = \epsilon$. Hence, $\lambda(\phi(E)) = 0$.

□

Theorem 12.60. *Let $R = \prod_{i=1}^{n}[a_i, b_i]$ be an n-rectangle. Let $T : R \to \mathbb{R}^n$ be the linear transformation defined by $T(\boldsymbol{e}_j) = \boldsymbol{e}_k$, $T(\boldsymbol{e}_k) = \boldsymbol{e}_j$ for some $j, k \in \{1, 2, 3, ..., n\}$, and $T(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for all $i \in \{1, 2, 3, ..., n\} \setminus \{j, k\}$. Then $Vol(T(R)) = Vol(R)$ and $T(R)$ is a rectangle.*

Proof. A point $(x_1, x_2, ..., x_n) \in R$ if and only if $(x_1, x_2, ..., x_{j-1}, x_k, x_{j+1}, ..., x_{k-1}, x_j, x_{k+1}, ..., x_n) \in T(R)$. Hence, $T(R) = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_{j-1}, b_{j-1}] \times [a_k, b_k] \times [a_{j+1}, b_{j+1}] ... \times [a_{k-1}, b_{k-1}] \times [a_j, b_j] \times ... \times [a_n, b_n]$. Thus, $Vol(T(R)) = \prod_{i=1}^{n}(b_i - a_i) = Vol(R)$ since switching the order of multiplication does not affect a product. □

Next, we observe that by using properties of determinants and elementary matrices we can show that the image of a rectangle under a linear transformation represented by multiplication by a matrix has a volume equal to its original volume times the determinant of the matrix (we also observe that translations don't change volumes since this will be useful later). This is done in a few steps.

Theorem 12.61. *Let $R = \prod_{i=1}^{n}[a_i, b_i]$ be an n-rectangle. Let $T : R \to \mathbb{R}^n$ be the linear transformation defined by $T(\boldsymbol{e}_j) = k\boldsymbol{e}_j$ for some non-zero constant k, and $T(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for $i \neq j$. Then $Vol(T(R)) = |k|Vol(R)$ and $T(R)$ is a rectangle.*

Proof. First, a point $(x_1, x_2, ..., x_j, ..., x_n) \in R$ if and only if $(x_1, x_2, ..., kx_j, ..., x_n) \in T(R)$. Hence, $T(R)$ is a rectangle whose edge factors are $[a_i, b_i]$ in all coordinates except for the jth coordinate, and whose jth coordinate is between ka_j and kb_j. If $k > 0$ then the jth edge factor is $[ka_j, kb_j]$ so that volume of $T(R)$ is $k\prod_{i=1}^{n}(b_i - a_i) = kVol(R) = |k|Vol(R)$. If $k < 0$ then the jth edge factor is $[kb_j, ka_j]$, so the volume is $Vol(T(R)) = \prod_{i=1}^{n}(b_i - a_i) = -kVol(R) = |k|Vol(R)$. □

Theorem 12.62. *Let $R = \prod_{i=1}^{n}[a_i, b_i]$ be an n-rectangle. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the translation $T(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{c}$, for some constant $\boldsymbol{c} = (c_1, c_2, c_3, ..., c_n) \in \mathbb{R}^n$. Then $Vol(T(R)) = Vol(R)$ and $T(R)$ is a rectangle.*

Proof. By definition, $T(R) = \prod_{i=1}^{n}[a_i + c_i, b_i + c_i]$ and $Vol(T(R)) = \prod_{i=1}^{n}(b_i + c_i - (a_i - c_i)) = \prod_{i=1}^{n}(b_i - a_i) = Vol(R)$. □

Theorem 12.63. *Let $R = \prod_{i=1}^{n}[a_i, b_i]$ be an n-rectangle. Let $t \in \mathbb{R}$ and let $T : R \to \mathbb{R}^n$ be the linear transformation defined by $T(\boldsymbol{e}_j) = \boldsymbol{e}_j + t\boldsymbol{e}_k$ for some $j, k \in \{1, 2, 3, ..., n\}$, and $T(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for all $i \in \{1, 2, 3, ..., n\} \setminus \{j\}$. Then $Vol(T(R)) = Vol(R)$.*

Proof. If $t = 0$ then this is the identity map, and the result follows. Assume that $t > 0$. Let $D = \prod_{\{1 \leq i \leq n | i \neq j\}} [a_i, b_i]$. Since T does not change any coordinate except for the jth coordinate on a point, all elements of $T(R)$ have coordinates other than their jth coordinate in D. Since $T(R) = \{(x_1, ..., x_j, ..., x_n) | (x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_n) \in D$ and $a_j + tx_k \leq x_j \leq b_j + tx_k\}$, $T(R)$ is a Jordan region by Theorem 12.53.

For $a_k \leq c < d \leq b_k$, we set $R_{[c,d]} = \{\mathbf{x} \in R | c \leq x_k \leq d\}$. Define $Q_{[c,d]} = \{\mathbf{x} \in \mathbb{R}^n | a_i \leq x_i \leq b_i$ if $i \in \{1, 2, ..., n\} \setminus \{j, k\}$, $c \leq x_k \leq d$ and $ct + a_j \leq x_j \leq dt + b_j\}$. Then $Q_{[c,d]}$ is a rectangle which contains $T(R_{[c,d]})$ since $ct + a_j$ is the smallest value of $x_j + tx_k$ and $dt + b_j$ is the largest value of $x_j + tx_k$ if $x_k \in [c, d]$ and $x_j \in [a_j, b_j]$.

If $d - c$ is small enough that $dt - ct < b_j - a_j$ then we also define $W_{[c,d]} = \{\mathbf{x} \in \mathbb{R}^n | a_i \leq x_i \leq b_i$ if $i \in \{1, 2, ..., n\} \setminus \{j, k\}$, $c \leq x_k \leq d$ and $dt + a_j \leq x_j \leq ct + b_j\}$. In this case, $W_{[c,d]} \subseteq T(R_{[c,d]})$ because for all $x_k \in [c, d]$ it is true that $ct \leq tx_k \leq dt$. Hence, if $dt + a_j \leq x_j \leq ct + bj$ then $a_j + tx_k \leq x_j \leq b_j + tx_k$.

Choose $m \in \mathbb{N}$ so that $\Delta_m = \dfrac{b_k - a_k}{m} < \dfrac{b_j - a_j}{t}$ so $\Delta_m t < b_j - a_j$. Let $P = \{a_k + i\Delta_m\}_{0 \leq i \leq m}$ be a partition of $[a_k, b_k]$. Then $\bigcup_{i=1}^{m} W_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]} \subseteq T(R) \subseteq \bigcup_{i=1}^{m} Q_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}$. Since $W_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]} \cap W_{[a_k+(i)\Delta_m, a_k+(i+1)\Delta_m]}$ consists of points whose x_k coordinate is $a_k + i\Delta m$, which is a subset of the boundaries of both rectangles (and volume zero) for each $1 \leq i \leq m - 1$, it follows that the collection $\{W_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}\}_{1 \leq i \leq m}$ of rectangles is non-overlapping. Likewise, $\{Q_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}\}_{1 \leq i \leq m}$ is a collection of non-overlapping rectangles. Similarly, the $\{R_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}\}_{1 \leq i \leq m}$ is a non-overlapping collection of rectangles which means that $\{T(R_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}\}_{1 \leq i \leq m}$ is a non-overlapping collection of Jordan regions by Theorem 12.58.

Hence, $\sum_{i=1}^{m} |W_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}| \leq Vol(T(R)) \leq \sum_{i=1}^{m} |Q_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}|$. The edge factors of the rectangles $Q_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}$ have lengths $b_i - a_i$ if $i \in \{1, 2, 3, ..., n\} \setminus \{j, k\}$. The kth edge factor is length Δ_m. The jth edge factor has length $b_j + (a_k + i\Delta m)t - ((a_k + (i-1)\Delta m)t + a_j) = (b_j - a_j) + t\Delta_m$. Thus, $\sum_{i=1}^{m} |Q_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}| = (\prod_{i \in \{1,2,...,n\} \setminus \{j,k\}} (b_i - a_i))(\sum_{i=1}^{m} \Delta_m)(b_j - a_j + t\Delta_m) = (b_j - a_j + t\Delta_m)(\prod_{i \in \{1,2,...,n\} \setminus \{j\}} (b_i - a_i))$ $= Vol(R) + t\Delta_m(\prod_{i \in \{1,2,...,n\} \setminus \{j\}} (b_i - a_i))$. Since we can make Δ_m as small as we wish by making m sufficiently large, it follows that $Vol(T(R)) \leq Vol(R)$.

The edge factors of the rectangles $W_{[a_k+(i-1)\Delta_m, a_k+i\Delta_m]}$ have lengths $b_i - a_i$ when $i \in \{1, 2, 3, ..., n\} \setminus \{j, k\}$. The kth edge factors has length Δ_m. The jth edge factor has length $(b_j - a_j) + ((a_k + (i-1)\Delta m)t - (a_k + i\Delta m)t = (b_j - a_j) - \Delta_m t$. Thus,

$\displaystyle\sum_{i=1}^{m}|W_{[a_k+(i-1)\Delta_m,a_k+i\Delta_m]}| = (\prod_{i\in\{1,2,...,n\}\setminus\{j,k\}}(b_i-a_i))(\sum_{i=1}^{m}\Delta_m)(b_j-a_j-t\Delta_m) = (b_j-a_j+t\Delta_m)(\prod_{i\in\{1,2,...,n\}\setminus\{j\}}(b_i-a_i)) = Vol(R)-t\Delta_m(\prod_{i\in\{1,2,...,n\}\setminus\{j\}}(b_i-a_i))$. Since we can make Δ_m as small as we wish by making m sufficiently large, it follows that $Vol(T(R)) \geq Vol(R)$ and so $Vol(T(R)) = Vol(R)$.

In the case where $t < 0$ the argument is similar.

□

Theorem 12.64. *Let $\phi : U \to \mathbb{R}^n$, where U is open in $\mathbb{R}^n$. Let $\overline{E} \subset U$ for some Jordan region E.*

(a) If ϕ is the translation $\phi(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{c}$ for some $\boldsymbol{c} \in \mathbb{R}^n$, or the linear transformation defined by $\phi(\boldsymbol{e}_j) = \boldsymbol{e}_k$ and $\phi(\boldsymbol{e}_k) = \boldsymbol{e}_j$ and $\phi(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for all other $i \in \{1, 2, 3, ..., n\}$, or the linear transformation defined by $\phi(\boldsymbol{e}_j) = \boldsymbol{e}_j + t\boldsymbol{e}_k$ for some $j, k \in \{1, 2, 3, ..., n\}$, and $\phi(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for all $i \in \{1, 2, 3, ..., n\} \setminus \{j\}$. Then $Vol(\phi(E)) = Vol(E)$.

(b) If ϕ is the linear transformation defined by $\phi(\boldsymbol{e}_j) = k\boldsymbol{e}_j$ for some non-zero constant k, and $\phi(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for $i \neq j$. Then $Vol(\phi(E)) = |k|Vol(E)$.

Proof. Note that $\phi(E)$ is a Jordan region by Theorem 12.58. Let $\epsilon > 0$.

(a) We know ϕ and ϕ^{-1} take rectangles to Jordan regions of the same volume by Theorems 12.60, 12.62, 12.63, and 12.58. Choose a finite collection of rectangles $\{R_i\}_{1\leq i\leq k}$ which covers E so that $\displaystyle\sum_{i=1}^{k}|R_i| < Vol(E) + \epsilon$. Then $\{\phi(R_i)\}_{1\leq i\leq k}$ is a cover of $\phi(E)$ with $\displaystyle\sum_{i=1}^{n}|\phi(R_i)| = \sum_{i=1}^{k}|R_i| < Vol(E) + \epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $Vol(\phi(E)) \leq Vol(E)$ by Theorem 12.24. Similarly, we can find a collection of rectangles $\{Q_i\}_{1\leq i\leq t}$ that covers $\phi(E)$ so that $\displaystyle\sum_{i=1}^{t}|Q_i| < Vol(\phi(E)) + \epsilon$. Hence, $\{\phi^{-1}(Q_i)\}_{1\leq i\leq t}$ covers E and $\displaystyle\sum_{i=1}^{t}|\phi^{-1}(Q_i)| =< Vol(\phi(E)) + \epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $Vol(E) \leq Vol(\phi(E))$. Hence, $Vol(\phi(E)) = Vol(E)$.

(b) By Theorem 12.61 we know that for any rectangle $R \subset U$, the image $\phi(R)$ is a rectangle and $Vol(\phi(R)) = |k||R|$. Note that $\phi^{-1}(\mathbf{e}_j) = \dfrac{1}{k}\mathbf{e}_j$, and $\phi^{-1}(\mathbf{e}_i) = \mathbf{e}_i$ for $i \neq j$. Thus, for any rectangle R it follows that $\phi^{-1}(R)$ is a rectangle and $Vol(\phi^{-1}(R)) = \dfrac{1}{|k|}|R|$.

Let $\{R_i\}_{1\leq i\leq k}$ be a collection of rectangles which covers E so that $\displaystyle\sum_{i=1}^{k}|R_i| < Vol(E) + \epsilon$. Then $\{\phi(R_i)\}_{1\leq i\leq k}$ is a cover of $\phi(E)$ with $\displaystyle\sum_{i=1}^{n}|\phi(R_i)| = \sum_{i=1}^{k}|k||R_i| < |k|Vol(E) + |k|\epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $Vol(\phi(E)) \leq |k|Vol(E)$.

Similarly, we can find a collection of rectangles $\{Q_i\}_{1\leq i\leq t}$ that covers $\phi(E)$ so that

$\sum_{i=1}^{t} |Q_i| < Vol(\phi(E)) + \epsilon$. Hence, $\{\phi^{-1}(Q_i)\}_{1 \leq i \leq t}$ is a collection of rectangles that covers E so that $\sum_{i=1}^{t} |\phi^{-1}(Q_i)| = \frac{1}{|k|} \sum_{i=1}^{t} |Q_i| < \frac{1}{|k|} Vol(\phi(E)) + \frac{1}{|k|}\epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $Vol(E) \leq \frac{1}{|k|} Vol(\phi(E))$ so $|k|Vol(E) \leq Vol(\phi(E)$. Hence, $Vol(\phi(E)) = |k|Vol(E)$. □

Theorem 12.65. *Let E be a Jordan region in $\mathbb{R}^n$, and let $T(\boldsymbol{x}) = A\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{R}^n$, where* $\det(A) \neq 0$. *Then $Vol(T(E)) = |\det(A)|Vol(E)$.*

Proof. Since $\det(A) \neq 0$ we can write $A = E_1 E_2 ... E_k$, where $E_1, E_2, ..., E_k$ are elementary matrices by Theorem 14.13 and for all $1 \leq i \leq k$, E_i is a matrix with corresponding linear transformation $T(\mathbf{x}) = E_i \mathbf{x}$ of one of the following forms: a matrix obtained by switching the ith row and jth row of the identity matrix (multiplication by which gives a linear transformation that interchanges the ith and jth standard basis vectors and takes all other standard basis vectors to themselves), a matrix obtained by adding the ith row times a number k to the jth row (multiplication by which is the linear transformation $T(\mathbf{e}_j) = \mathbf{e}_j + k\mathbf{e}_i$), or a matrix multiplying the jth row of the identity matrix by the non-zero number k, multiplication by which gives the transformation defined by $T(\mathbf{e}_j) = k\mathbf{e}_j$ and $T(\mathbf{e}_i) = \mathbf{e}_i$ for $i \neq j$.

Applying Theorem 12.64 k times, we see $Vol(T(E)) = |E_1||E_2||E_3|...|E_k|Vol(E)$. This is equal to $|\det(A)|Vol(E)$ since the product of the determinants of matrices is the determinant of the product of those matrices by Theorem 14.13. □

The following is a development with aspects paralleling that found in Buck's Advanced Calculus text with some parts paralleling part of the approach found in Wade's Advanced Calculus, though it is also significantly different in many respects. Despite the differences in the approach, a student would probably benefit from looking at the proofs for Change of Variables given in those two texts in order to improve context and perspective to the development below.

Definition 97

Let U be an open set in $\mathbb{R}^n$ and let $F : \mathcal{J} \to \mathbb{R}$, where $\mathcal{J}$ is the set of all Jordan regions whose closures are contained in U. Then we say that $\lim_{R \downarrow \mathbf{p}} F(R) = L$ if for every $\epsilon > 0$ there is a $\delta > 0$ so that if R is a cube in U containing $\mathbf{p}$ and $diam(R) < \delta$ then $|F(R) - L| < \epsilon$. We say that F is *additive* if $F(E_1 \cup E_2) = F(E_1) + F(E_2)$ whenever E_1 and E_1 are disjoint. We say that F is *monotone* if for any pair of Jordan regions E_1, E_2 so that $E_1 \subseteq E_2$ it is true that $F(E_1) \leq F(E_2)$.

If $f(\mathbf{p}) = \lim_{R \downarrow \mathbf{p}} F(R)$ exists for all $\mathbf{p} \in D$ for some set $D \subseteq \mathbb{R}^n$ then we say that $\lim_{R \downarrow \mathbf{p}} F(R) = f(\mathbf{p})$ *uniformly* on D if for every $\epsilon > 0$ there is some $\delta > 0$ so that if R is a cube containing $\mathbf{p}$ and $diam(R) < \delta$ then $|F(R) - f(\mathbf{p})| < \epsilon$.

We say that F is *differentiable* on D if $F'(\mathbf{p}) = \lim_{R \downarrow \mathbf{p}} \frac{F(R)}{Vol(R)}$ exists at each $\mathbf{p} \in D$. We say that F is *uniformly differentiable* if this limit exists uniformly on D.

We say that F is *volume continuous* if $F(E) = 0$ for every Jordan region $E \subseteq U$ so that $Vol(E) = 0$.

We may refer to functions like F described above as set functions to indicate to the reader that their domains are sets of points rather than individual points in $\mathbb{R}^n$.

Theorem 12.66. *Let $f : U \to \mathbb{R}$ be continuous, where U is an open subset of $\mathbb{R}^n$, and let $\mathcal{J}$ be the set of all Jordan regions contained in U. Let $F : \mathcal{J} \to \mathbb{R}$ be defined by $F(E) = \int_E f$. Then:*

(a) F is additive and volume continuous with $F'(\boldsymbol{p}) = f(\boldsymbol{p})$ for each $\boldsymbol{p} \in U$.

(b) F is uniformly differentiable on any compact set $K \subset U$.

(c) If f is non-negative then F is monotone.

Proof. (a) We know F is additive by Theorem 12.44. Since $\int_E f = 0$ if $Vol(E) = 0$, F is also volume continuous.

(b) Let $\mathbf{p} \in U$. Given any $\epsilon > 0$ we can find $\delta > 0$ so that if $|\mathbf{p} - \mathbf{x}| < \delta$ then $|f(\mathbf{x}) - f(\mathbf{p})| < \epsilon$, so if R is a cube containing $\mathbf{p}$ and $diam(R) < \delta$ then $f(\mathbf{p}) - \epsilon < f(\mathbf{x}) < f(\mathbf{p}) + \epsilon$ for all $\mathbf{x} \in R$. Thus, $(f(\mathbf{p}) - \epsilon)Vol(R) \leq \int_R f \leq (f(\mathbf{p}) + \epsilon)Vol(R)$ by Theorem 12.44. Hence, $f(\mathbf{p}) - \epsilon \leq \frac{F(R)}{Vol(R)} \leq f(\mathbf{p}) + \epsilon$, which means that $f(\mathbf{p}) = \lim_{R \downarrow \mathbf{p}} \frac{F(R)}{Vol(R)}$.

On a compact set $K \subset U$ we can (by the Lebesgue Number Lemma) pick $\delta_1 > 0$ so that any set of diameter less than δ_1 which intersects K is a subset of U. Since K is compact, f is uniformly continuous, which means that we can make the choice of $0 < \delta < \delta_1$ as above so that if $|\mathbf{p} - \mathbf{x}| < \delta$ then $|f(\mathbf{x}) - f(\mathbf{p})| < \epsilon$ for all $\mathbf{p} \in K$, and therefore $f(\mathbf{p}) - \epsilon \leq \frac{F(R)}{Vol(R)} \leq f(\mathbf{p}) + \epsilon$ for all $\mathbf{p} \in K$ and cubes R so that $diam(R) < \delta$, which means that F is uniformly differentiable on K.

(c) If f is non-negative then if E_1, E_2 are Jordan regions with $E_1 \subseteq E_2 \subseteq U$ then it follows that $\int_{E_2 \setminus E_1} f \geq 0$, so $\int_{E_2} f = \int_{E_1} f + \int_{E_2 \setminus E_1} f \geq \int_{E_1} f$.

□

The following is a little like an analogue for the fundamental theorem of calculus for functions on sets as defined above.

Theorem 12.67. *Let $F : \mathcal{J} \to \mathbb{R}$ be additive and volume continuous, where $\mathcal{J}$ is the set of all Jordan regions in an open set $U \subseteq \mathbb{R}^n$, so that F is differentiable on U and uniformly differentiable on compact convex subsets of U, with $F'(\boldsymbol{p}) = f(\boldsymbol{p})$ for each $\boldsymbol{p} \in U$. Then f*

is continuous on U and $F(Q) = \int_Q f$ for every cube $Q \subset U$. If F is non-negative on some open set V so that $\overline{V} \subset U$ then $F(E) = \int_E f$ for every Jordan region $E \subseteq V$.

Proof. We first show that f is continuous. Let $\epsilon > 0$. Let $\mathbf{p} \in U$. Then $\mathbf{p} \in \overline{C_r(\mathbf{p})} \subset U$ for some $r > 0$. Since $\overline{C_r(\mathbf{p})}$ is compact and convex, F is uniformly differentiable on $\overline{C_r(\mathbf{p})}$. Hence, we can choose $\delta_1 > 0$ so that if R is a cube with $diam(R) < \delta_1$ and $\mathbf{x} \in R \cap \overline{C_r(\mathbf{p})}$ then $|\frac{F(R)}{Vol(R)} - f(\mathbf{x})| < \frac{\epsilon}{2}$. We can choose a cube W centered at $\mathbf{p}$ with diameter less than δ_1 containing a ball $B_\gamma(\mathbf{p})$ for some $\gamma > 0$. If $|\mathbf{p} - \mathbf{x}| < \gamma$ then $\mathbf{x}, \mathbf{p} \in W$, so $|f(\mathbf{p}) - f(\mathbf{x})| \leq |f(\mathbf{p}) - \frac{F(W)}{Vol(W)}| + |\frac{F(W)}{Vol(W)} - f(\mathbf{x})| < \epsilon$. Hence, f is continuous at each point $\mathbf{p} \in U$.

We define a function $G : \mathcal{J} \to \mathbb{R}$ by $G(E) = \int_E f$. By Theorem 12.66, we know that G is differentiable on U and uniformly differentiable on compact subsets of U, with $F'(\mathbf{p}) = G'(\mathbf{p}) = f(\mathbf{p})$ for all $\mathbf{p} \in U$.

Let Q be a cube contained in U. Then Q is compact and convex so we can find a $\delta > 0$ so that if R is a cube containing a point $\mathbf{p}$ of Q and $diam(R) < \delta$ then $|f(\mathbf{p}) - \frac{F(R)}{Vol(R)}| < \frac{\epsilon}{2}$ and $|f(\mathbf{p}) - \frac{G(R)}{Vol(R)}| < \frac{\epsilon}{2}$, so $|\frac{F(R)}{Vol(R)} - \frac{G(R)}{Vol(R)}| < \epsilon$, so $|F(R) - G(R)| < \epsilon Vol(R)$.

Let $G = \{R_i\}_{1 \leq i \leq k}$ be a grid on Q consisting of cubes, with $|G| < \delta$. Let $S = \bigcup_{i=1}^{k} \partial(R_i)$. Then $Vol(S) = 0$ and since F is additive and volume continuous we know that $F(Q) = F(\bigcup_{i=1}^{n} R_i) = F(S) + \sum_{i=1}^{k} F(R_i^\circ) = \sum_{i=1}^{k} F(R_i^\circ) = \sum_{i=1}^{k} F(R_i)$. By definition, $G(Q) = \int_Q f = \sum_{i=1}^{k} G(R_i)$. Thus, $F(Q) - G(Q) < \epsilon \sum_{i=1}^{k} Vol(R_i) = \epsilon Vol(Q)$. Since this is true for all $\epsilon > 0$ it follows that $F(Q) = G(Q)$.

Next, assume that F is non-negative on V and $E \subseteq V \subseteq \overline{V} \subseteq U$. Suppose that $f(\mathbf{p}) < 0$ for some $\mathbf{p} \in V$ and choose $t > 0$ so that if $|\mathbf{x} - \mathbf{p}| < t$ then $|f(\mathbf{x}) - f(\mathbf{p})| < \frac{|f(\mathbf{p})|}{2}$, so $f(\mathbf{x}) < \frac{f(\mathbf{p})}{2}$. Thus, if Q is a cube containing $\mathbf{p}$ which is contained in $B_t(\mathbf{p})$ then $F(Q) = \int_Q f \leq \frac{f(\mathbf{p})}{2}|Q| < 0$, which is impossible. We conclude that f is also non-negative. Hence, by Theorem 12.66, we know that F is monotone.

Let E be a Jordan region. By Theorem 12.49 we can find an $\eta > 0$ so that any grid on a cube containing E whose mesh is less than η has the property that all upper and lower sums, inner sums and outer sums are within a distance ϵ of the integral of f. We can also chose η so that any set of diameter less than η intersecting $\overline{E}$ is a subset of U. Let $H = \{Q_i\}_{1 \leq i \leq m}$ be a grid consisting entirely of cubes on a cube containing E with $|H| < \eta$. Since f is non-negative, $0 \leq m_i \leq M_i$ for each $1 \leq i \leq m$.

Since F is monotone, it follows that $F(\bigcup I(E,H)) \leq F(E) \leq F(\bigcup O(E,H))$ since $\bigcup I(E,H) \subseteq E \subseteq \bigcup O(E,H))$. Hence, $\int_E f - \epsilon < L(f,H)^\circ = \sum_{Q_i \in I(E,H)} m_i Vol(Q_i) \leq$

$\displaystyle\sum_{Q_i\in I(E,H)}\int_{Q_i} f = \sum_{Q_i\in I(E,H)} F(Q_i) = F(\bigcup I(E,H)) \le F(E) \le F(\bigcup O(E,H)) = \sum_{Q_i\in O(E,H)} F(Q_i) = \sum_{Q_i\in O(E,H)}\int_{Q_i} f \le \sum_{Q_i\in O(E,H)} M_i Vol(Q_i) = \overline{U}(f,H) < \int_E f + \epsilon$. Thus, we see that $|F(E) - \displaystyle\int_E f| < \epsilon$ for all $\epsilon > 0$, so $\displaystyle\int_E f = F(E)$.

□

Theorem 12.68. *Let $\phi : U \to \mathbb{R}^n$ be C^1, one to one and have $\Delta_\phi \neq 0$ on U. Let $C_{2r}(\boldsymbol{z}) \subset U$ and let $0 < s < \frac{1}{2}$. Let $|\phi(\boldsymbol{x}) - \boldsymbol{x}| < sr$ for all $\boldsymbol{x} \in C_{2r}(\boldsymbol{z})$. Then $C_{2r(1-s)}(\boldsymbol{z}) \subseteq \phi(C_{2r}(\boldsymbol{z}))$.*

Proof. Let $\mathbf{q} \in C_{2r(1-s)}(\mathbf{z})$. Set $\psi(\mathbf{x}) = |\phi(\mathbf{x}) - \mathbf{q}|$ on $C_{2r}(\mathbf{z})$. Since ψ is continuous it takes on a minimum value m on $C_{2r}(\mathbf{z})$. Note that $|\phi(\mathbf{q}) - \mathbf{q}| < sr$ so $m < sr$. For any point $\mathbf{y} \in \partial(C_{2r}(\mathbf{z}))$. Then $y_j = z_j \pm r$ for some $j \in \{1,2,3,...,n\}$. Since $|\phi(\mathbf{y}) - \mathbf{y}| < sr$ we know $|\phi(\mathbf{y})_j - \mathbf{y}_j| < sr$ so either $\phi(\mathbf{y})_j > z_j + r - sr$ or $\phi(\mathbf{y})_j < z_j - r + sr$, which means that $|y_j - z_j| > (1-s)r$. Since $(1-s)r > sr$ (because $s < \frac{1}{2}$), it follows that $\psi(\mathbf{y}) > m$ and therefore ψ cannot take on a minimum on the $\partial(C_{2r}(\mathbf{z}))$. This means that there is some $\mathbf{p} \in C_{2r}(\mathbf{z})^\circ$ so that ψ takes on a minimum at $\mathbf{p}$.

Let $u_1, u_2, ..., u_n$ be the variables for ϕ and $x_1, x_2, ..., x_n$ be the variables for $g(\mathbf{x}) = \displaystyle\sum_{i=1}^{n}(x_i - q_i)^2$. Because $\psi^2 = g(\phi(\mathbf{x}))$ takes on a local minimum at $\mathbf{p}$, it follows that all partial derivatives of ψ^2 with respect to variables $u_1, u_2, ..., u_n$ are zero at $\mathbf{p}$. Thus,

$$2(x_1 - q_1)\frac{\partial x_1}{\partial u_1}(\mathbf{p}) + 2(x_2 - q_2)\frac{\partial x_2}{\partial u_1}(\mathbf{p}) + ... + 2(x_n - q_n)\frac{\partial x_n}{\partial u_1}(\mathbf{p}) = 0$$
$$2(x_1 - q_1)\frac{\partial x_1}{\partial u_2}(\mathbf{p}) + 2(x_2 - q_2)\frac{\partial x_2}{\partial u_2}(\mathbf{p}) + ... + 2(x_n - q_n)\frac{\partial x_n}{\partial u_2}(\mathbf{p}) = 0$$
$$\vdots$$
$$2(x_1 - q_1)\frac{\partial x_1}{\partial u_n}(\mathbf{p}) + 2(x_2 - q_2)\frac{\partial x_2}{\partial u_n}(\mathbf{p}) + ... + 2(x_n - q_n)\frac{\partial x_n}{\partial u_n}(\mathbf{p}) = 0.$$

Since $\Delta_\phi(\mathbf{p}) \neq 0$ this system has a unique solution, namely $x_i = q_i$ for each i, which means that $\phi(\mathbf{p}) = (x_1, x_2, ..., x_n) = \mathbf{q}$ and therefore $C_{2r(1-s)}(\mathbf{z}) \subseteq \phi(C_{2r}(\mathbf{z}))$.

□

For the next theorem, we will want to distinguish between functions from smaller dimensions fixing certain coordinates and functions from spaces of larger dimensions. Similar to notation used previously, we will use the notation $(\mathbf{x}, \mathbf{y})$ represents the vector whose first n coordinates are the coordinates of $\mathbf{x}$ and whose last n coordinates are those of $\mathbf{y}$).

Theorem 12.69. *(a) Let $\epsilon_1, \epsilon_2 > 0$. If $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^n$ and $|\boldsymbol{x} - \boldsymbol{w}| < \epsilon_1$ and $|\boldsymbol{y} - \boldsymbol{z}| < \epsilon_2$ then $|(\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{w}, \boldsymbol{z})| < \sqrt{\epsilon_1^2 + \epsilon_2^2}$.*

(b) Let U, V be open sets in $\mathbb{R}^n$. Then $U \times V$ is open in $\mathbb{R}^{2n}$.

(c) Let A, B be closed in $\mathbb{R}^n$. Then $A \times B$ is closed in $\mathbb{R}^{2n}$.

(d) Let K, M be a compact subsets of $\mathbb{R}^n$. Then $K \times M$ is a compact subset of $\mathbb{R}^{2n}$.

Proof. (a) Since $\sum_{i=1}^{n}(x_i - w_i)^2 < \epsilon_1^2$ and $\sum_{i=1}^{n}(y_i - z_i)^2 < \epsilon_2^2$, $\sum_{i=1}^{n}(x_i - w_i)^2 + \sum_{i=1}^{n}(y_i - z_i)^2 < \epsilon_1^2 + \epsilon_2^2$, so $|(\mathbf{x}, \mathbf{y}) - (\mathbf{w}, \mathbf{z})| < \sqrt{\epsilon_1^2 + \epsilon_2^2}$.

(b) If U or V is empty then the cross product $U \times V$ is empty and therefore open. Otherwise, let $(\mathbf{x}, \mathbf{y}) \in U \times V$. Then $\mathbf{x} \in U$ and $\mathbf{y} \in V$. Hence, we can find $\epsilon_{\mathbf{x}}, \epsilon_{\mathbf{y}} > 0$ so that $B_{\epsilon_{\mathbf{x}}}(\mathbf{x}) \subset U$ and $B_{\epsilon_{\mathbf{y}}}(\mathbf{y}) \subset V$. Let $m = \min\{\epsilon_{\mathbf{x}}, \epsilon_{\mathbf{y}}\}$. Let $(\mathbf{w}, \mathbf{z}) \in B_m(\mathbf{x}, \mathbf{y})$. Since $|(\mathbf{x}, \mathbf{y}) - (\mathbf{w}, \mathbf{z})|$ is greater than or equal to both $|\mathbf{x} - \mathbf{w}|$ and $|\mathbf{y} - \mathbf{z}|$, we know that $|\mathbf{x} - \mathbf{w}| < m$ so $\mathbf{w} \in B_{\epsilon_{\mathbf{x}}}(\mathbf{x})$, and similarly, $|\mathbf{y} - \mathbf{z}| < m$ so $\mathbf{z} \in B_{\epsilon_{\mathbf{y}}}(\mathbf{y})$. Therefore $B_m(\mathbf{x}, \mathbf{y}) \subseteq (B_{\epsilon_{\mathbf{x}}}(\mathbf{x}) \times B_{\epsilon_{\mathbf{y}}}(\mathbf{y})) \subseteq U \times V$ which is open.

(c) If A, B are empty then their product is empty and therefore closed. Otherwise, note that since A, B are closed $\mathbb{R}^n \setminus A$ is open and $\mathbb{R}^n \setminus B$ is open. By part (b) we know that $(\mathbb{R}^n \setminus A) \times \mathbb{R}^n$ and $\mathbb{R}^n \times (\mathbb{R}^n \setminus B)$ are open in $\mathbb{R}^{2n}$, which means that $W = ((\mathbb{R}^n \setminus A) \times \mathbb{R}^n) \cup (\mathbb{R}^n \times (\mathbb{R}^n \setminus B))$ is also open. Note $(\mathbf{x}, \mathbf{y}) \notin A \times B$ if and only if $\mathbf{x} \notin A$ or $\mathbf{y} \notin B$. If $\mathbf{x} \in \mathbb{R}^n \setminus A$ then $(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^n \setminus A) \times \mathbb{R}^n$ and if $\mathbf{y} \notin B$ then $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times (\mathbb{R}^n \setminus B)$. Thus, $\mathbb{R}^{2n} \setminus (A \times B) = W$ is open, and therefore $A \times B$ is closed.

(d) By the Heine-Borel Theorem, K and M are compact if and only if they are closed and bounded. Since K and M are bounded we can find numbers $r, s > 0$ so that $K \subset B_r(\mathbf{0})$, and $M \subset B_s(\mathbf{0})$. Then by part (a) we know $K \times M \subset B_{\sqrt{r^2+s^2}}(\mathbf{0})$ in $\mathbb{R}^{2n}$ which means that $K \times M$ is bounded.

By part (c), we know that $K \times M$ is closed. Hence, by the Heine-Borel Theorem we can conclude that $K \times M$ is compact.

□

Theorem 12.70. *Let $\boldsymbol{p} \in U$, an open set in $\mathbb{R}^n$. Let $\phi : U \to \mathbb{R}^n$ be one to one and C^1 with $\Delta_\phi \neq 0$ on U. Then $\lim_{R \downarrow \boldsymbol{p}} \frac{Vol(\phi(R))}{Vol(R)} = |\Delta_\phi(\boldsymbol{p})|$. Furthermore, if K is a convex compact subset of U then $\lim_{R \downarrow \boldsymbol{p}} \frac{Vol(\phi(R))}{Vol(R)} = |\Delta_\phi(\boldsymbol{p})|$ uniformly on K.*

Proof. Let K be a compact convex subset of U and let $0 < s < \frac{1}{2}$. Let $\epsilon = \frac{s}{\sqrt{n}}$

For each $\mathbf{p} \in J$ we define $\psi_{\mathbf{p}}(\mathbf{x}) = (D\phi(\mathbf{p}))^{-1}\phi(\mathbf{x})$. Then ψ is C^1 and $D\psi_{\mathbf{p}}(\mathbf{p}) = I$. Define $g_{\mathbf{p}}(\mathbf{x}) = \psi(\mathbf{x}) - \psi(\mathbf{p}) + \mathbf{p}$, and define $f_{\mathbf{p}}(\mathbf{x}) = g_{\mathbf{p}}(\mathbf{x}) - \mathbf{x}$. Let $h_{\mathbf{p}}(\mathbf{x}) : U \to \mathbb{R}$ be defined by $h_{\mathbf{p}}(\mathbf{x}) = \sqrt{n} \sum_{i=1}^{m} \sum_{j=1}^{n} |\frac{\partial f_{\mathbf{p}_i}}{\partial x_j}(\mathbf{x})|$.

For each of these functions $\psi_{\mathbf{p}}, g_{\mathbf{p}}, f_{\mathbf{p}}, h_{\mathbf{p}} : U \to \mathbb{R}^n$. We define corresponding functions $\psi(\mathbf{p}, \mathbf{x}), g(\mathbf{p}, \mathbf{x}), f(\mathbf{p}, \mathbf{x}) : U \times U \to \mathbb{R}^n$ and $h(\mathbf{p}, \mathbf{x}) : U \times U \to \mathbb{R}$ defined by $\psi(\mathbf{p}, \mathbf{x}) = \psi_{\mathbf{p}}(\mathbf{x})$, $g(\mathbf{p}, \mathbf{x}) = g_{\mathbf{p}}(\mathbf{x})$, $f(\mathbf{p}, \mathbf{x}) = f_{\mathbf{p}}(\mathbf{x})$ and $h(\mathbf{p}, \mathbf{x}) = h_{\mathbf{p}}(\mathbf{x})$. Note that $Df_{\mathbf{p}}(\mathbf{p}) = \mathbf{0}_{n \times n}$ and $f_{\mathbf{p}}(\mathbf{p}) = \mathbf{0}$ for each $\mathbf{p} \in \mathbb{R}^n$. These functions are all continuous since the partial derivatives of ψ are continuous, which means that, in particular $h(\mathbf{p}, \mathbf{x})$ is uniformly continuous on the set $K \times K$, which is compact by Theorem 12.69. Thus, we can find a number $\delta_1 > 0$ so that if $|(\mathbf{y}, \mathbf{z}) - (\mathbf{p}, \mathbf{x})| < \delta_1$ then $|h(\mathbf{y}, \mathbf{z}) - h(\mathbf{p}, \mathbf{x})| < \epsilon$. More specifically, for any $\mathbf{p} \in K$ it is true that if $|\mathbf{x} - \mathbf{p}| < \delta_1$ then $|h_{\mathbf{p}}(\mathbf{p}) - h_{\mathbf{p}}(\mathbf{x})| < \epsilon$. Since $\mathbf{h}_{\mathbf{p}}(\mathbf{p}) = 0$ this means that $\mathbf{h}_{\mathbf{p}}(\mathbf{x}) < \epsilon$ for all $\mathbf{x} \in K$ so that $|\mathbf{x} - \mathbf{p}| < \delta_1$.

Fix some $\mathbf{p} \in K$. Let $C_r(\mathbf{q})$ be a cube of side length less than $\delta = \frac{\delta_1}{\sqrt{n}}$ which contains

$\mathbf{p}$. Then $C_r(\mathbf{q}) \subset B_{\delta_1}(\mathbf{p})$ which means that $h(\mathbf{x}) < \epsilon$ on $C_r(\mathbf{q})$. If $\mathbf{x}, \mathbf{y} \in C_r(\mathbf{q})$ then $L(\mathbf{x}, \mathbf{y}) \subset C_r(\mathbf{q})$ since cubes are convex, which means that $|f_{\mathbf{p}}(\mathbf{p}) - f_{\mathbf{p}}(\mathbf{x})| = |f_{\mathbf{p}}(\mathbf{x})| \leq (\max\limits_{\mathbf{w} \in C_r(\mathbf{q})} h_{\mathbf{p}}(\mathbf{w}))|\mathbf{p} - \mathbf{x}|$ by Theorem 11.18. This is less than $\epsilon|\mathbf{p} - \mathbf{x}| \leq \epsilon(diam(C_r(\mathbf{q}))) = \epsilon\sqrt{n}r = \dfrac{s}{\sqrt{n}}\sqrt{n}r = sr$.

This means that $|g_{\mathbf{p}}(\mathbf{x}) - \mathbf{x}| < sr$ for all $\mathbf{x} \in C_r(\mathbf{q})$. Hence, by Theorem 12.68, we know that $g_{\mathbf{p}}(C_r(\mathbf{q}))$ contains a cube $C_{r(1-s)}(\mathbf{q})$. We also know that $\psi(C_r(\mathbf{q}))$ is a Jordan region by Theorem 12.58. Hence, $Vol(C_{r(1-s)}(\mathbf{q})) \leq Vol(g_{\mathbf{p}}(C_r(\mathbf{q})))$. Translations preserve volumes on sets by Theorems 12.62 and 12.64, so $Vol(g_{\mathbf{p}}(C_{r(1-s)}(\mathbf{q}))) = Vol(\psi(C_{r(1-s)}(\mathbf{q})))$, so we also know that $Vol(C_{r(1-s)}(\mathbf{q})) \leq Vol(\psi(C_r(\mathbf{q})))$.

If $\mathbf{y} \in C_r(\mathbf{q})$ then $q_j - \dfrac{r}{2} \leq y_j \leq q_j + \dfrac{r}{2}$ for each $j \in \{1, 2, 3, ..., n\}$. Since $|g_{\mathbf{p}}(\mathbf{y}) - \mathbf{y}| < sr$ we know $|g_{\mathbf{p}}(\mathbf{y})_j - \mathbf{y}_j| < sr$ so $q_j - \dfrac{r}{2} - sr \leq y_j - sr < g_{\mathbf{p}}(\mathbf{y})_j < y_j + sr \leq q_j + \dfrac{r}{2} + sr$. Hence, it follows that $g_{\mathbf{p}}(C_r(\mathbf{q})) \subset C_{r(1+2s)}(\mathbf{q})$. Thus, $Vol(g_{\mathbf{p}}(C_r(\mathbf{q}))) \leq Vol(C_{r(1+2s)}(\mathbf{q}))$, so $Vol(\psi(C_r(\mathbf{q}))) \leq Vol(C_{r(1+2s)}(\mathbf{q}))$. Hence, if $C_r(\mathbf{q})$ is a cube whose diameter is less than δ which contains $\mathbf{p}$ then $\dfrac{Vol(C_{r(1-s)}(\mathbf{q}))}{Vol(C_r(\mathbf{q}))} \leq \dfrac{Vol(\psi(C_r(\mathbf{q}))}{Vol(C_r(\mathbf{q}))} \leq \dfrac{Vol(C_{r(1+2s)}(\mathbf{q}))}{Vol(C_r(\mathbf{q}))}$. Note that the left and right inequalities can be made as close as we wish to one by making s small enough since $Vol(C_{r(1-s)}(\mathbf{q})) = (r(1-s))^n$ and $Vol(C_{r(1+2s)}(\mathbf{q})) = (r(1+2s))^n$, which are both continuous functions of s, so $\lim\limits_{R \downarrow \mathbf{p}} \dfrac{Vol(\psi(R))}{Vol(R)} = 1$. Since the choice of δ was independent of the choice of $\mathbf{p} \in K$ it is also true that $\lim\limits_{R \downarrow \mathbf{p}} \dfrac{Vol(\psi(R))}{Vol(R)} = 1$ uniformly on K.

Since $|\Delta_\phi(\mathbf{x})|$ is continuous on the compact set K, by the Extreme Value Theorem we can find an $M > 0$ so that $M > \max\limits_K |\Delta_\phi(\mathbf{x})|$. Let $\epsilon_1 > 0$. Since $\lim\limits_{R \downarrow \mathbf{p}} \dfrac{Vol(\psi(R))}{Vol(R)} = 1$ uniformly on K, we can choose $\delta > 0$ so that for any $\mathbf{p} \in K$, if R is a cube of diameter less than δ and $\mathbf{p} \in R$ then $1 - \dfrac{\epsilon_1}{M} < \dfrac{Vol(\psi(R))}{Vol(R)} < 1 + \dfrac{\epsilon_1}{M}$. Since $\psi(\mathbf{x}) = (D\phi(\mathbf{p}))^{-1}\phi(\mathbf{x})$, by Theorem 12.65 $Vol(\psi(R)) = Vol((D\phi(\mathbf{p}))^{-1}\phi(R)) = |\det(D\phi(\mathbf{p}))^{-1}|Vol(\phi(R))$. We also know that $\det(D\phi(\mathbf{p}))^{-1} = \dfrac{1}{\Delta_\phi(\mathbf{p})}$.

Thus, $1 - \dfrac{\epsilon_1}{|\Delta_\phi(\mathbf{p})|} < 1 - \dfrac{\epsilon_1}{M} < \dfrac{1}{|\Delta_\phi(\mathbf{p})|}\dfrac{Vol(\phi(R))}{Vol(R)} < 1 + \dfrac{\epsilon_1}{M} < 1 + \dfrac{\epsilon_1}{|\Delta_\phi(\mathbf{p})|}$, which means that $|\Delta_\phi(\mathbf{p})| - \epsilon_1 < \dfrac{Vol(\phi(R))}{Vol(R)} < |\Delta_\phi(\mathbf{p})| + \epsilon_1$, and therefore $\lim\limits_{R \downarrow \mathbf{p}} \dfrac{Vol(\phi(R))}{Vol(R)} = |\Delta_\phi(\mathbf{p})|$ uniformly (since the choice of δ depended only on M and K and not on $\mathbf{p}$) on every compact convex subset of U.

□

Theorem 12.71. *Let U be an open set in $\mathbb{R}^n$ and let $\phi : U \to \mathbb{R}^n$ be one to one and C^1 with $\Delta_\phi \neq 0$ on U. Then there is a volume continuous, additive, monotone set function $F : \mathcal{J} \to [0, \infty)$, where $\mathcal{J}$ is the set of all Jordan regions whose closures are contained in U so that $F(C) = Vol(\phi(C))$ for every cube $C \subset U$.*

Such a function F must satisfy $F(E) = Vol(\phi(E))$ for every Jordan region E whose closure is contained in U.

Proof. First, define $F(E) = \displaystyle\int_{\phi(E)} 1 = Vol(E)$ for every Jordan region E whose closure is contained in U. Then $F(C) = Vol(C)$ for every cube $C \subset U$, and F is positive and monotone. Since $Vol(\phi(E)) = 0$ if E is a Jordan region of volume zero whose closure is contained in U by Theorem 12.59, F is volume continuous. Hence, such an F exists.

Next, assume that $F : \mathcal{J} \to [0, \infty)$, where $\mathcal{J}$ is the set of all Jordan regions whose closures are contained in U, is a set function so that $F(C) = Vol(\phi(C))$ for every cube $C \subset U$. Let E be a Jordan region whose closure is contained in U. By the Lebesgue Number Lemma we can find a $\delta_1 > 0$ so that if S is a set of diameter less than or equal to δ_1 which intersects $\overline{E}$ then $S \subset U$. By Theorem 12.56, we can find an $m > 0$ and a $\delta_2 > 0$ so that if a cube C has diameter less than δ_2 and intersects $\overline{E}$ then $\phi(C)$ is contained in a cube of volume less than $mVol(C)$.

By Theorem 12.46 can find $\delta_3 > 0$ so that if H is a grid on a rectangle containing E with $|H| < \delta_3$ then $V(\partial(E), H) < \dfrac{\epsilon}{m}$.

Choose a grid $G = \{R_i\}_{1 \le i \le k}$ on a cube containing E so that $|G| < \min\{\delta_1, \delta_2, \delta_3\}$. Then $\displaystyle\bigcup_{R_i \in I(E,G)} \phi(R_i) \subseteq \phi(E)$, so $\displaystyle\sum_{R_i \in I(E,G)} Vol(\phi(R_i)) = F(\bigcup I(E,G)) \le F(E)$ since F is monotone. Since $\displaystyle\bigcup_{R_i \in O(E,G)} \phi(R_i) \supseteq \phi(E)$, we know $\displaystyle\sum_{R_i \in O(E,G)} Vol(\phi(R_i)) = F(\bigcup O(E,G)) \ge$ $F(E)$. Also, we know that $\displaystyle\sum_{R_i \in O(E,G)} Vol(\phi(R_i)) - \sum_{R_i \in I(E,G)} Vol(\phi(R_i)) = \sum_{R_i \in S(\partial(E),G)} Vol(\phi(R_i)) \le$ $\displaystyle m \sum_{R_i \in S(\partial(E),G)} Vol(R_i) < \frac{\epsilon}{m} m = \epsilon$. Thus, $Vol(\phi(E)), F(E) \in [\displaystyle\sum_{R_i \in I(E,G)} Vol(\phi(R_i)), \sum_{R_i \in O(E,G)} Vol(\phi(R_i))]$ and hence $|Vol(\phi(E)) - F(E)| < \epsilon$. Since this is true for all $\epsilon > 0$ it follows that $F(E) = Vol(\phi(E))$. □

Theorem 12.72. *Let U be an open set in $\mathbb{R}^n$ and let $\phi : U \to \mathbb{R}^n$ be one to one and C^1 with $\Delta_\phi \neq 0$ on U. Let E be a Jordan region so that $\overline{E} \subset U$. Then $Vol(\phi(E)) = \displaystyle\int_{\phi(E)} 1 =$ $\displaystyle\int_E |\Delta_\phi|$.*

Proof. Let $F : \mathcal{J} \to [0, \infty)$, where $\mathcal{J}$ is the set of all Jordan regions whose closures are contained in U, be a volume continuous, additive, monotone set function so that $F(C) = Vol(\phi(C))$ for every cube $C \subseteq U$. Then F is additive and volume continuous and monotone. We know $\displaystyle\int_{\phi(E)} 1 = Vol(\phi(E))$ by Theorem 12.44. We know that $F(E) = Vol(\phi(E))$ for every Jordan region E whose closure is contained in U (and that such a function F exists) by Theorem 12.71.

By Theorem 12.70, F is uniformly differentiable on any compact subset of U with derivative $F'(\mathbf{p}) = \Delta_\phi(\mathbf{p})$. Hence, by Theorem 12.67 for every Jordan E region whose closure is contained in U, $F(E) = \displaystyle\int_E |\Delta_\phi| = Vol(\phi(E)) = \int_{\phi(E)} 1$. □

Theorem 12.73. *Change of Variables (or Transformation of Variables). Let $\overline{E} \subset U$, where E is a Jordan region and U is open in $\mathbb{R}^n$. Let $\phi : U \to \mathbb{R}^n$ be a one to one* *continuously differentiable function so that $\Delta_\phi \neq 0$ on U, and let f be integrable on $\phi(E)$. Then $\displaystyle\int_{\phi(E)} f = \int_E f \circ \phi|\Delta_\phi|$.*

Proof. First, we show that $f \circ \phi$ is integrable on E. Let D_f be the set of discontinuities of f. Then $\lambda(D_f) = 0$ by the Lebesgue Characterization of Riemann Integrability. Since ϕ^{-1} a one to one continuously differentiable function so that $\Delta_{\phi^{-1}} \neq 0$ on $\phi(U)$ by the Inverse Function Theorem, it follows from Theorem 12.59 that $\lambda(\phi^{-1}(D_f)) = 0$. For any point $\mathbf{p} \in E \setminus \phi^{-1}(D_f)$ we know that ϕ is continuous at $\mathbf{p}$ and f is continuous at $\phi(\mathbf{p})$ since $\phi(\mathbf{p}) \notin D_f$ because ϕ is one to one, and since $|\Delta_\phi|$ is also continuous we know that $f \circ \phi|\Delta_\phi|$ is continuous at $\mathbf{p}$. Hence, if $D_{f\circ\phi|\Delta_\phi|}$ is the set of discontinuities of $f \circ \phi|\Delta_\phi|$ in E then $\lambda(D_{f\circ\phi|\Delta_\phi|}) = 0$ so $f \circ \phi|\Delta_\phi|$ is integrable on E. Note that for a rectangle R in interior of the range of ϕ, since ϕ^{-1} is also a one to one continuously differentiable function so that $\Delta_{\phi^{-1}} \neq 0$ on the open set $\phi(U)$ by the Inverse Function Theorem, so by Theorem 12.72 it follows that $\displaystyle\int_{\phi^{-1}(R)} |\Delta_\phi| = \int_1 R = Vol(R)$.

Next, observe that $f = (\frac{f+|f|}{2}) - (\frac{|f|-f}{2})$, where both $\frac{f+|f|}{2}$ and $\frac{|f|-f}{2}$ are non-negative and integrable. Hence, if we can prove the theorem for non-negative functions then this will prove the result for $\frac{f+|f|}{2}$ and $\frac{|f|-f}{2}$, which would show $\displaystyle\int_{\phi(E)} \frac{f+|f|}{2} = \int_E (\frac{f+|f|}{2}) \circ \phi|\Delta_\phi|$ and $\displaystyle\int_{\phi(E)} \frac{|f|-f}{2} = \int_E (\frac{|f|-f}{2}) \circ \phi|\Delta_\phi|$. From this, by subtracting the second from the first of these integrals, we would have $\displaystyle\int_{\phi(E)} f = \int_E f \circ \phi|\Delta_\phi|$ (even if f takes on negative values). Thus, it is sufficient to prove the theorem for non-negative functions f.

We assume that f is non-negative. Let $\epsilon > 0$. Choose $\delta_1 > 0$ so that if S is a set whose diameter is no more than δ_1 and S intersects $\overline{\phi(E)}$ then $R_i \subset \phi(U)$. This is possible (by the Lebesgue Number Lemma) since $\phi(\overline{E}) = \overline{\phi(E)}$ is compact and contained in the open set $\phi(U)$. We can also choose $\delta_2 > 0$ so that if H is a grid on a rectangle containing $\phi(E)$ with $|H| < \delta_2$ then all upper sums, upper inner sums, upper outer sums, lower sums, lower inner sums and lower outer sums of $\phi(E)$ with respect to grid H are within ϵ of $\displaystyle\int_{\phi(E)} f$ by Theorem 12.49. Let $\delta = \min\{\delta_1, \delta_2\}$.

Let $\{G_i\}_{1\le i\le k}$ be a grid on a rectangle containing $\phi(E)$ so that $|G| < \delta$. Let $V = \displaystyle\bigcup_{R_i\in S(\phi(E),G)} \phi^{-1}(R_i)$. By Theorem 12.72, we know that $\displaystyle\int_{\phi^{-1}(R_i)} |\Delta_\phi| = \int_{R_i} 1 = Vol(R_i)$ for any $R_i \in G$. So, $U(f,G) = \displaystyle\sum_{R_i\in S(\phi(E),G)} M_i Vol(R_i) = \sum_{R_i\in S(\phi(E),G)} M_i \int_{\phi^{-1}(R_i)} |\Delta_\phi| \geq \sum_{R_i\in S(\phi(E),G)} \int_{\phi^{-1}(R_i)} f \circ \phi|\Delta_\phi| = \int_V f \circ \phi|\Delta_\phi|$. Also, $L(f,G)^\circ = \displaystyle\sum_{R_i\in I(\phi(E),G)} m_i Vol(R_i) = \sum_{R_i\in I(\phi(E),G)} m_i \int_{\phi^{-1}(R_i)} |\Delta_\phi| \leq \sum_{R_i\in I(\phi(E),G)} \int_{\phi^{-1}(R_i)} f \circ \phi|\Delta_\phi| = \int_W f \circ \phi|\Delta_\phi|$, where $W = \displaystyle\bigcup_{R_i\in I(\phi(E),G)} \phi^{-1}(R_i)$.

We know $L(f,G)^\circ \leq L(f,G)$ since f is non-negative, and that $L(f,G) \leq \int_{\phi(E)} f \leq U(f,G)$ and we know that $U(f,G) - L(f,G)^\circ < 2\epsilon$. Since $E \subseteq V$ we know that $\int_V f \circ \phi|\Delta_\phi| = \int_E f \circ \phi|\Delta_\phi|$, so $\int_E f \circ \phi|\Delta_\phi| \leq U(f,G)$. Since $W \subseteq E$ and f is non-negative it follows that $L(f,G)^\circ \leq \int_W f \circ \phi|\Delta_\phi| \leq \int_E f \circ \phi|\Delta_\phi|$. Thus we see $\int_E f \circ \phi|\Delta_\phi|, \int_{\phi(E)} f \in [L(f,G)^\circ, U(f,G)]$, so $|\int_{\phi(E)} f - \int_E f \circ \phi|\Delta_\phi|| < 2\epsilon$. Since this is true for all $\epsilon > 0$ we conclude that $\int_{\phi(E)} f = \int_E f \circ \phi|\Delta_\phi|$. This completes the proof.

□

We often want to use polar coordinates to integrate over circles in the plane, or an analogue of polar coordinates in three dimensions to deal with triple integrals. The first of these is cylindrical coordinates, which is convenient for integrating over cylinders or prisms over circular sectors, and which is also helpful for many other shapes like cones and paraboloids. In cylindrical coordinates we just add a third variable z which is equal to the z coordinate of the original point. Thus, if the projection of a point $((x,y,z)$ onto the xy-plane can be represented as $(x,y,0)$ and in the plane the point $(x,y) = (r,\theta)$ in polar coordinates using the principal polar represenatation of (x,y), then the cylindrical coordinates representation for (x,y,z) is (r,θ,z). Hence, using the usual representations in polar coordinates, we have that $x = x\cos(\theta), y = r\sin(\theta), z = z$ is the one to one (except at $r = 0$ and possibly at $\theta = 2\pi$ if we map the entire closed cylinder) C^1 map from the rectangle $[0,R] \times [0,2\pi] \times [-h,h]$ to the cylinder from $z = -h$ to $z = h$ over a disk of radius R centered at the origin. Hence, if a region E is enclosed within such a cylinder (which any bounded region is) then E is the image of some region D in the $r\theta z$ three dimensional space under the cylindrical coordinates mapping thus defined. Hence $\int\int\int_E f(x,y,z)dV = \int\int\int_D f(r\cos(\theta), r\sin(\theta), z)|\det J|dV$, where $J = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) & 0 \\ \sin(\theta) & r\cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Thus, expanding along the third row we get $|\det(J)| = r$. Hence, the conversion formula from rectangular to cylindrical coordinates is:

$$\int\int\int_E f(x,y,z)dV = \int\int\int_D f(r\cos(\theta), r\sin(\theta), z)rdV$$

Thus, the Jacobian factor by which we integrate when we change variables in cylindrical coordinates is the same as the factor we multiply by when converting a two variable integral to polar coordinates. As with polar coordinates, this can be intuitively described by looking at a small rectangle with side lengths $\Delta r, \Delta\theta, \Delta z$ at point (r_i, θ_j, z_k) in the $r\theta z$ space, and noticing that this corresponds to a section of a cylinder in the xyz space where the base of this solid has area approximately $r_i\Delta\theta\Delta r$ (just as with the polar coordinates derivation if the change in θ is small then the base is approximately rectangular and the length of one side is Δr and the length of the other is roughly $r_i\Delta\theta$ since it is a circular section). The height of this section of cylinder is just Δz, so the volume of the resulting cylindrical section is approximately $r_i\Delta\theta\Delta r\Delta z$. Thus, we would expect $\lim\limits_{n\to\infty}\lim\limits_{m\to\infty}\lim\limits_{t\to\infty}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{t} f(r_i\cos(\theta_j), r_i\sin(\theta_j), z_k)r_i\Delta\theta\Delta r\Delta z$ to approach the integral of f over the original corresponding region that is mapped onto by the rectangles in the sums listed, so we would anticipate that these integrals would be equal geometrically.

Example 12.11. *Find the integral* $\displaystyle\int\int\int_E x^2 dV$*, where* E *is the the half of the cone bounded by* $z = \sqrt{x^2+y^2}$ *and* $z = 4$ *where* $y \geq 0$*.*

Solution. Vertically oriented cones with points at the origin tend to work well with conversion to cylindrical coordinates. In this case, the intersection of the cone and the plane listed is along $\sqrt{x^2+y^2} = 4$, a radius four circle centered at the origin. The portion of that projection with non-negative y value is a semicircle (making the angle $\theta \leq \pi$). For the z coordinate over that projection, the smallest value of z would be at $\sqrt{x^2+y^2} = r$ and the largest value is four. So, describing these points in cylindrical coordinates in the bounds of an integral we would have $\displaystyle\int_0^4\int_0^{\pi}\int_r^4 r^2\cos^2(\theta)r dz d\theta dr$. This is $\displaystyle 2\int_0^4\int_0^{\frac{\pi}{2}} zr^3\cos^2(\theta)\Big|_r^4 d\theta dr =$
$\displaystyle 2\int_0^4\int_0^{\frac{\pi}{2}} 4r^3\cos^2(\theta)d\theta dr - 2\int_0^4 r^3 dr\int_0^{\frac{\pi}{2}} r^4\cos^2(\theta)d\theta dr = 8\int_0^4 r^3 dr\int_0^{\frac{\pi}{2}}\cos^2(\theta)d\theta - 2\int_0^4 r^4 dr\int_0^{\frac{\pi}{2}}\cos^2(\theta)d\theta$
$\displaystyle = 8(\frac{4^4}{4})(\frac{1}{2})(\frac{\pi}{2}) - 2(\frac{4^5}{5})(\frac{1}{2})(\frac{\pi}{2}) = 128\pi - \frac{512\pi}{5} = \frac{128\pi}{5}.$

□

Another way to generalize polar coordinates to three coordinate systems is spherical coordinates. This system, as the name suggests, offers a way to more easily integrate over spheres, regions between spheres, and sections of spheres. Some such sections look more like ice cream cones with spheres at the top, depending on the bounds of the variables. For this coordinate system we let $\rho = \sqrt{x^2+y^2+x^2}$, which is the distance from the origin to the point (x, y, z) in space. We then let θ be the same angle as in cylindrical coordinates (which is the same θ as polar coordinates for the projected point (x, y) in the plane). Finally, we let ϕ be the (smallest) angle from the positive z-axis to the line segment from the origin to (x, y, z). We notice then, using trigonometry, that $z = \rho\cos(\phi)$. Likewise, we see that $r = \rho\sin(\phi)$ (where r is the same as it is in cylindrical coordinates, the distance from the origin to the point (x, y) in the plane). Thus, $x = \rho\sin(\phi)\cos(\theta)$ and $y = \rho\sin(\phi)\sin(\theta)$. This allows us to express the transformation from the $\rho, \theta\phi$ space to the xyz space using $x = \rho\sin(\phi)\cos(\theta)$ and $y = \rho\sin(\phi)\sin(\theta)$ and $z = \rho\cos(\phi)$. Using the transformation of variables formula, we get:

$$\int\int\int_E f(x,y,z)dV = \int\int\int_D f(\rho\sin(\phi)\cos(\theta), \rho\sin(\phi)\sin(\theta), \rho\cos(\phi)|\det J|dV$$

The matrix $J = \begin{bmatrix} \sin(\phi)\cos(\theta) & \sin(\phi)\sin(\theta) & \cos(\phi) \\ -\rho\sin(\phi)\sin(\theta) & \rho\sin(\phi)\cos(\theta) & 0 \\ \rho\cos(\phi)\cos(\theta) & \rho\cos(\phi)\sin(\theta) & -\rho\sin(\phi) \end{bmatrix}$. Expanding along the third column, we get $-\rho^2\cos(\phi)[\sin^2(\theta)\sin(\phi)\cos(\phi) + \cos^2(\theta)\sin(\phi)\cos(\phi)]$

$-\rho^2\sin(\phi)[\sin^2(\phi)\cos^2(\theta)+\sin^2(\phi)\sin^2(\theta)]$
$= -\rho^2\cos^2(\phi)\sin(\phi)-\rho^2\sin^3(\phi) = -\rho^2\sin(\phi)(\cos^2(\phi)+\sin^2(\phi)) = -\rho^2\sin(\phi)$. This means that $|\det J| = \rho^2\sin(\phi)$ (assuming we do not use values of ϕ that are not in the $[0,\pi]$ interval, which is unnecessary).

Thus, the formula for converting to spherical coordinates is:

$$\int\int\int_E f(x,y,z)dV = \int\int\int_D f(\rho\sin(\phi)\cos(\theta),\rho\sin(\phi)\sin(\theta),\rho\cos(\phi)\rho^2\sin(\phi)dV$$

As with cylindrical coordinates, an intuitive interpretation can be seen if we consider the spherical wedge formed by taking the image of a small rectangle in the $\rho\theta\phi$ space under the spherical coordinates transformation at $(\rho_i,\theta_j,\phi_k$ we get an image which is approximately a rectangle and we notice that the length of edge of the wedge section corresponding to the change in the angle ϕ is approximately $\rho\Delta\phi$ throughout the wedge section (thought it is slightly longer on the outer edge than the inner edge), and the length of the edge along the circular section corresponding to a change $\Delta\theta$ is $r\Delta\theta$ which is $\rho\sin(\phi)\Delta\theta$, and that the depth of the wedge section is $\Delta\rho$, so the volume of the corresponding wedge section is approximately $\rho^2\sin(\phi)\Delta\rho\Delta\theta\Delta\phi$ (it approaches this value as the number of subdivisions becomes large) . Hence, the integral should correspond to the limit:

$$\lim_{n\to\infty}\lim_{m\to\infty}\lim_{r\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} f(\rho_i\sin(\phi_k)\cos(\theta_j),\rho_i\sin(\phi_k)\sin(\theta_j),\rho_i\cos(\phi_k))\rho_i^2\sin(\phi_k)\Delta\rho\Delta\theta\Delta\phi.$$

Attempting to decide which coordinate system to use to evaluate an integral is not always straightforward and frequently multiple choices will work well. While the following rule of thumb is not true in general, it is often the case that it is usually not worth converting a planar integral to polar coordinates unless the region to be integrated over is circular sector or a region between two circular sectors (unless it is easy to see that the planar region is another nice planar curve in polar coordinates like a rose). For the most part, it there isn't much point to converting to cylindrical coordinates unless the projection of the region onto one of the coordinate planes is a nice polar region of the types we just described. We typically don't want to convert to spherical coordinates unless the outer surface of the region is on a sphere (and the solid is preferably a region within a spherical wedge or between two spherical wedges), in which case spherical coordinates is often better than cylindrical coordinates. In other integrals where the projection onto the coordinate axes is a nice polar region (such as cones with flat tops, cylinders or paraboloids), cylindrical coordinates tend to work better than spherical coordinates. If none of these criteria apply, you are probably better off leaving the integral in rectangular coordinates.

Example 12.12. *Find* $\int_{-2}^{2}\int_{0}^{\sqrt{4-x^2}}\int_{0}^{\sqrt{4-x^2-y^2}} x^2y^2z^2dzdydx.$

Solution. We see that the region where $-2\leq x\leq 2$ and $0\leq y\leq\sqrt{9-x^2}$ is the upper half of the radius two disk centered at the origin. Then, if $0\leq z\leq\sqrt{4-x^2-y^2}$ we end up with the solid over that half disk inside the radius two sphere about the origin. Since this is a portion of a sphere, we would guess that spherical coordinates would probably give the nicest transformation. The spherical bounds would have angles θ ranging from zero to π, as normal for the upper half of a disk. For each angle of θ the angles ϕ from the positive

z-axis range from zero to $\frac{\pi}{2}$ (coming halfway down to the negative z-axis). The radii at each of these angles ranges from zero to two because in a radius two sphere. This gives us $\int_0^{\pi}\int_0^{\frac{\pi}{2}}\int_0^2 \rho^2\sin^2(\phi)\cos^2(\theta)\rho^2\sin^2(\phi)\sin^2(\theta)\rho^2\cos^2(\phi)\rho^2\sin(\phi)d\rho d\phi d\theta =$
$\int_0^{\pi}\cos^2(\theta)\sin^2(\theta)d\theta\int_0^{\frac{\pi}{2}}\sin^5(\phi)\cos^2(\phi)d\phi\int_0^2\rho^8 d\rho = (2)(\frac{1}{(4)(2)})(\frac{\pi}{2})(\frac{(4)(2)(1)}{(7)(5)(3)(1)})(\frac{2^9}{9})$
$= \frac{512\pi}{945}$.

□

Using the fact that converting to spherical simplifies integration over a sphere, it becomes easier to use transformation of variables to integrate over regions bounded by ellipsoids just as such transformations made it easier to integrate over elliptic disks once we knew polar transformations.

Example 12.13. *Let E be the region bounded by the ellipsoid $\frac{x^2}{4}+\frac{y^2}{9}+\frac{z^2}{16}=1$. Find $\int\int\int_E z^4 dV$.*

Solution. We set $x = 2u, y = 3v$ and $z = 4w$ to get $u^2+v^2+w^2 = 1$, the unit sphere, meaning that this transformation takes the unit sphere onto this ellipsoid. We find the Jacobian $\det\begin{bmatrix}2&0&0\\0&3&0\\0&0&4\end{bmatrix} = 24$. Thus, if E is the unit ball, the integral becomes $\int\int\int_E 24(4w)^4 dV =$
$(24)(256)\int_0^{2\pi}\int_0^{\pi}\int_0^1(\rho\cos(\phi))^4\rho^2\sin(\phi)d\rho d\phi d\theta = (24)(256)(2\pi)\int_0^{\pi}\cos^4(\phi)\sin(\phi)d\phi\int_0^1\rho^6 d\rho =$
$(24)(256)(2\pi)(2)(\frac{(3)(1)}{(5)(3)(1)})\frac{1}{7} = \frac{24576\pi}{35}$.

□

Spherical and cylindrical and polar coordinates with respect to different axes:

Sometimes we want to use polar coordinates over a coordinate plane yz that we want to project onto. This can be done easily by just selecting a variable to correspond to the new x and y (assign $y = r\cos(\theta)$ and $z = r\sin(\theta)$ for instance) and then converting the two variable integral to polar coordinates and doing the integral as normal.

Sometimes we want to use cylindrical coordinates over a solid E which is easier to project onto another plane (say the yz-plane again). We would then let $x = x$, $y = r\cos(\theta)$ and $z = r\sin(\theta)$. It would also have been find to set $z = r\cos(\theta)$ and $y = r\sin(\theta)$. We just have to be consistent about replacing the corresponding variable with its transformed form in the integrand.

Sometimes we want to use spherical coordinates and measure the angle from a different direction for the ϕ angle. Again, let's say we would like ϕ to be measured from the positive x-axis, so that $\rho = \sqrt{x^2+y^2+z^2}$ again, but now $x = \rho\cos(\phi)$, $y = \rho\sin(\phi)\sin(\theta)$, $z = \rho\sin(\phi)\cos(\theta)$. Or, we could have switched that z and y were equal to that would have worked just as well.

In each case, however, we would have to make sure that the integral bounds for describing the region correspond to whatever variable definitions we assign.

Frequently, it is easier to think of this procedure as just interchanging two variables everywhere and keeping the variable definitions the same. This alternates the position of the solid and switches the variables corresponding to the switched axes in the integrand. In this manner, we don't have to re-define the transformation and instead switch the variables themselves.

Example 12.14. *Find the mass of the region E bounded by $x = y^2 + z^2$ and $x = 1$ so that $z \geq 0$, with density function $\rho(x, y, z) = y^2 x$.*

Solution. This is better done with cylindrical coordinates than spherical coordinates. The projection of this solid onto the yz-plane is the upper half of a unit disk since $y^2 + z^2 = 1$ at the intersection of the two surfaces.

Let's proceed by several approaches. First, we will use a rotated version of cylindrical coordinates. We will have $x = x$ and set $y = r\cos(\theta)$ and $z = r\sin(\theta)$ where $r = \sqrt{y^2 + z^2}$, and θ is the angle made with the positive y axis measured towards the positive z-axis (counterclockwise as viewed from the positive x-axis looking at the yz-plane). The disk would be traced out over $0 \leq \theta \leq \pi$ and $0 \leq r \leq 1$. Over that disk the function is $x = y^2 + z^2$, so $r^2 \leq x \leq 1$. Thus, the integral is $\int_0^\pi \int_0^1 \int_{r^2}^1 xr^3 \cos^2(\theta) dxdrd\theta = \int_0^\pi \int_0^1 r^3 \frac{x^2}{2} \cos^2(\theta)\Big|_{r^2}^1 drd\theta = \frac{1}{2}\int_0^\pi \int_0^1 r^3 \cos^2(\theta) - r^5 \cos^2(\theta) drd\theta = \frac{1}{2}\int_0^\pi \cos^2(\theta)d\theta \int_0^1 r^3 - r^5 drd\theta = (\frac{1}{2})(2)(\frac{1}{2})(\frac{\pi}{2})(\frac{r^4}{4} - \frac{r^6}{6})\Big|_0^1 = \frac{\pi}{48}$.

Next, let's use the approach where we re-orient the axes (switching z and z). So, the mass listed would be the same as the mass if we took the region E bounded by $z = x^2 + y^2$ and $z = 1$ so that $x \geq 0$, with density function $\rho(x, y, z) = y^2 z$. Because we changed the variables in both the integrand and the description of the region, we should end up with the same mass as the original solid had. Projecting down onto the xy-plane we have the right half of the unit disk. This gives an integral $\int_{-\frac{pi}{2}}^{\frac{\pi}{2}} \int_0^1 \int_{r^2}^1 zr^2 \sin^2(\theta) dzdrd\theta$. The only difference between this integral and the previous one is the bounds on the first integral, but we see that $\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin^2(\theta)d\theta = \int_0^\pi \cos^2(\theta)d\theta$ (so the integral will give the same answer).

Finally, what if we tried to change the cylindrical coordinate system so that $x = x$ and set $y = r\sin(\theta)$ and $z = r\cos(\theta)$ where $r = \sqrt{y^2 + z^2}$ and θ is measured as the previous transformation of variables listed above. Then the bounds for the region would have $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$ and the integral would be $\int_{-\frac{pi}{2}}^{\frac{\pi}{2}} \int_0^1 \int_{r^2}^1 xr^2 \sin^2(\theta) dxdrd\theta$, which has the same value as the integral in the previous method.

□

So, the bounds change based on our assignment but as long as we remain consistent and describe the region in the new variables we have some flexibility about how we reassign cylindrical coordinates.

Spherical coordinates are similar. Here is an example.

Example 12.15. *Find* $\int_E ydV$ *if* E *is the solid bounded by* $x^2+y^2+z^2=4$ *and* $y=\sqrt{x^2+z^2}$.

Solution. We can see that E is a spherical wedge which is symmetric about the positive y-axis. In spherical coordinates, it would be easier to describe this wedge if ϕ were measured from the positive y-axis instead of the z-axis.

First, let's do the problem by reassigning the spherical coordinates transformation on the variables. We would measure ϕ as the angle from the positive y-axis to the line segment connecting the origin to each point. We would have $y=\rho\cos(\phi)$ and $x=\rho\sin(\phi)\cos(\theta)$ and $z=\rho\sin(\phi)\sin(\theta)$ (we could have reversed the assignments of x and z and this would have altered how θ was measured, but as long as our description in the bounds was consistent with the assignment it would have been fine).

With this assignment, we see that when $\theta=0$ we are starting where $z=0$ and x is positive. When $\theta=\frac{\pi}{2}$ x would be zero and z would be positive, so we can see that θ is being measured from the positive x direction towards the positive z-axis direction, which is clockwise as viewed from the positive y-axis. The angle ϕ is zero at the y-axis and $\frac{\pi}{4}$ when $y=r$. Having observed this, we would be able to describe the solid listed as $\int_0^{2\pi}\int_0^{\frac{\pi}{4}}\int_0^2 \rho\cos(\phi)\rho^2\sin(\phi)d\rho d\phi d\theta = \int_0^{2\pi} 1d\theta \int_0^{\frac{\pi}{4}}(\frac{1}{2})\sin(2\phi)d\phi \int_0^2 \rho^3 d\rho = (2\pi)(-\frac{1}{4}\cos(2\phi)\Big|_0^{\frac{\pi}{4}})(\frac{\rho^4}{4}\Big|_0^2) = 2\pi$.

As an alternate approach, we will switch y and z variables so that the spherical wedge is symmetric about the positive z axis which makes spherical coordinates in the original system convenient to use. We would then see the region E as being the solid bounded by $x^2+y^2+z^2=4$ and $z=\sqrt{x^2+y^2}$. This would have $0\le\theta\le 2\pi$ and $0\le\phi\le\frac{\pi}{4}$ and $0\le\rho\le 2$. The integrand y would be replaced by z since we are switching those two variables. This changes the integral to $\int_0^{2\pi}\int_0^{\frac{\pi}{4}}\int_0^2 \rho\cos(\phi)\rho^2\sin(\phi)d\rho d\phi d\theta$, which is the same as before.

□

Exercises:

Exercise 12.1. *Let $f : E \to \mathbb{R}$ be integrable on a Jordan region E in $\mathbb{R}^n$ so that for some $i \in \{1, 2, 3, ..., n\}$, if $(x_1, x_2, ..., x_i, ..., x_n) \in E$ then $(x_1, x_2, ..., -x_i, ..., x_n) \in E$, and $f(x_1, x_2, ..., x_i, ..., x_n) = -f(x_1, x_2, ..., -x_i, ..., x_n)$. Prove that $\displaystyle\int_E f = 0$.*

Exercise 12.2. *Give an example of a bounded function f which is bounded on a rectangle R that is not integrable on R. Prove f is not integrable.*

Exercise 12.3. *Let $f, g : R \to \mathbb{R}$ be integrable, where R is a rectangle in $\mathbb{R}^n$. Prove that $h(\boldsymbol{x}) = \max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$ is also integrable on R.*

Exercise 12.4. *Let $f : R \to \mathbb{R}$ be continuous, where R is a rectangle in $\mathbb{R}^m$, and let $g(E) \subseteq R$, where $g : E \to \mathbb{R}^m$ is integrable and E is a Jordan region in $\mathbb{R}^n$. Then prove that $f \circ g$ is integrable on E.*

Exercise 12.5. *Give an example of an integrable function f on a rectangle R in $\mathbb{R}^n$ so that the set of discontinuities of f is a dense subset of R.*

Exercise 12.6. *Show that for any natural numbers n and m it is possible to find $f : R \to \mathbb{R}$ which is integrable on a rectangle R in $\mathbb{R}^n$ and a function $g : E \to R$ which is continuous so that E is a Jordan region in $\mathbb{R}^m$ and $f \circ g$ is not integrable on E.*

Exercise 12.7. *Use Change of Variables to integrate the function $f(x, y) = x^2$ over*

(a) The region E enclosed by the ellipse $\dfrac{x^2}{4} + \dfrac{y^2}{9} = 1$ and

(b) The region E bounded by the lines $y - x = 0$, $y - x = 4$, $3x + y = 0$, $3x + y = 2$.

Exercise 12.8. *Let E be a Jordan region in $\mathbb{R}^n$. Prove or disprove: E° is also a Jordan region.*

Exercise 12.9. *Prove that any bounded subset E of $\mathbb{R}^n$ which has only finitely many limit points has volume zero.*

Exercise 12.10. *Prove that every open ball (and every closed ball) in $\mathbb{R}^n$ is a Jordan region.*

Exercise 12.11. *Show that the countable union of Jordan regions need not be a Jordan region, even if the union is bounded.*

Exercise 12.12. *Let E be a Jordan region in $\mathbb{R}^n$. Define the dilation of E by a factor of $k > 0$ to be $kE = \{k\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x} \in E\}$. Prove that kE is a Jordan region and $Vol(kE) = k^n Vol(E)$.*

Exercise 12.13. *Prove the Mean Value Theorem for Multiple Integrals. Let $f, g : E \to \mathbb{R}$ be integrable with $g(\boldsymbol{x}) \geq 0$ on the Jordan region E in $\mathbb{R}^n$. Then there is a number $c \in [\inf f(E), \sup f(E)]$ so that $c \displaystyle\int_E g = \int_E fg$. If f is continuous and E is connected then there is a value $\boldsymbol{w} \in E$ so that $f(\boldsymbol{w}) = c$. In particular, there is a number $c \in [\inf f(E), \sup f(E)]$ so that $cVol(E) = \displaystyle\int_E f$.*

Exercise 12.14. *Let $f : E \to \mathbb{R}$ be integrable on the Jordan region E in $\mathbb{R}^n$, where f is continuous at the point $\boldsymbol{z} \in E^\circ$. Prove that $\displaystyle\lim_{r \to 0^+} \frac{1}{Vol(B_r(\boldsymbol{z}))} \int_{B_r(\boldsymbol{z})} f = f(\boldsymbol{z})$.*

Exercise 12.15. *Let $f, g : E \to \mathbb{R}$ be integrable on the Jordan region E in $\mathbb{R}^n$. Show that if $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for every point $\boldsymbol{x} \in \mathbb{Q}^n \cap E$ then $\displaystyle\int_E f = \int_E g$.*

Solutions:

Solution to Exercise 12.1. *Let $f : E \to \mathbb{R}$ be integrable on a Jordan region E in $\mathbb{R}^n$ so that for some $j \in \{1, 2, 3, ..., n\}$, if $(x_1, x_2, ..., x_j, ..., x_n) \in E$ then $(x_1, x_2, ..., -x_j, ..., x_n) \in E$, and $f(x_1, x_2, ..., x_j, ..., x_n) = -f(x_1, x_2, ..., -x_j, ..., x_n)$. Prove that $\int_E f = 0$.*

Proof. Since E is a Jordan region, E is bounded so we can find $r > 0$ so that $E \subset R = \prod_{i=1}^{n}[-r, r]$. Let $\epsilon > 0$. Since E is integrable, we can find a $k \in \mathbb{N}$ so that if we set $P_i = \{-r, -r + \frac{2r}{k}, -r + \frac{4r}{k}, ..., r\}$ for all $1 \leq i \leq n$, then the grid G induced by $P_1, P_2, ..., P_n$ has small enough mesh so that $|\int_E f - \sum_{R_i \in G} f(\mathbf{x}_i^*)|R_i|| < \epsilon$ regardless of choice of marking points $\mathbf{x}_i^* \in R_i$. Choose the mid points of each rectangle $R_i \in G$ as the marking points x_i^*. Let H_P be the set of rectangles in G containing points with positive jth component, and let H_N be the set of rectangles in G containing points with negative jth component. For each $R_i \in H$, let $R_{(i,N)}$ be the element of H_N which consists of the points of R_i with jth component negated. Then if $\mathbf{x}_i^*$ is the mid point of H and $\mathbf{x}_{(i,N)}^*$ is the midpoint of $R_{(i,N)}$ we know that $f(\mathbf{x}_i^*) = -f(\mathbf{x}_{(i,N)}^*)$.

Thus, $\sum_{R_i \in G} f(\mathbf{x}_i^*)|R_i| = \sum_{R_i \in H} f(\mathbf{x}_i^*)|R_i| + \sum_{R_{(i,N)} \in H_N} f(\mathbf{x}_{(i,N)}^*)|R_{(i,N)}| = 0$. Thus, $|\int_E f| < \epsilon$ for every $\epsilon > 0$ and so $\int_E f = 0$.

□

Solution to Exercise 12.2. *Give an example of a bounded function f which is bounded on a rectangle R that is not integrable on R. Prove f is not integrable.*

Proof. Let R be any rectangle in $\mathbb{R}^n$. Let $f(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathbb{Q}^n$ and let $f(\mathbf{x}) = 1$ otherwise. Then for any grid G on R we have $U(f, G) = |R|$ and $L(f, G) = 0$, so $U(f, G) - L(f, G) = |R| > 0$, which means that f is not integrable.

□

Solution to Exercise 12.3. *Let $f, g : R \to \mathbb{R}$ be integrable, where R is a rectangle in $\mathbb{R}^n$. Prove that $h(\boldsymbol{x}) = \max\{f(\boldsymbol{x}), g(\boldsymbol{x})\}$ is also integrable on R.*

Proof. Let $\epsilon > 0$. Choose a grid $G = \{R_i\}_{i=1}^k$ on R so that $U(f, G) - L(f, G) < \frac{\epsilon}{2}$ and $U(g, G) - L(g, G) < \frac{\epsilon}{2}$. If $M_i(f) = \sup_{R_i} f(\mathbf{x})$ and $M_i(g) = \sup_{R_i} g(\mathbf{x})$ and $M_i = \sup_{R_i} h(\mathbf{x})$ then $M_i = \max\{M_i(f), M_i(g)\}$. Likewise, if $m_i(f) = \inf_{R_i} f(\mathbf{x})$ and $m_i(g) = \inf_{R_i} g(\mathbf{x})$ and $m_i = \inf_{R_i} h(\mathbf{x})$ then $m_i = \max\{m_i(f), m_i(g)\}$. If $M_i(f) \geq M_i(g)$ and $m_i(f) \geq m_i(g)$ then $M_i - m_i = M_i(f) - m_i(f)$. If $M_i(g) \geq M_i(f)$ and $m_i(g) \geq m_i(f)$ then $M_i - m_i = M_i(g) -$

$m_i(g)$. If $M_i(f) \geq M_i(g)$ and $m_i(f) \leq m_i(g)$ then $M_i - m_i = M_i(f) - m_i(g) \leq M_i(f) - m_i(f)$. If $M_i(f) \leq M_i(g)$ and $m_i(f) \geq m_i(g)$ then $M_i - m_i = M_i(g) - m_i(f) \leq M_i(g) - m_i(g)$.

Thus, $U(h,G) - L(h,G) = \sum\limits_{R_i \in G} (M_i - m_i)|R_i| \leq \sum\limits_{R_i \in G} \max\{M_i(f) - m_i(f), M_i(g) - m_i(g)\}|R_i| \leq \sum\limits_{R_i \in G} (M_i(f) - m_i(f))|R_i| + \sum\limits_{R_i \in G} (M_i(g) - m_i(g))|R_i| = U(f,G) - L(f,G) + U(g,G) - L(g,G) < \epsilon$. Hence, h is integrable on R. □

Solution to Exercise 12.4. *Let $f : R \to \mathbb{R}$ be continuous, where R is a rectangle in $\mathbb{R}^m$, and let $g(E) \subseteq R$, where $g : E \to \mathbb{R}^m$ is integrable and E is a Jordan region in $\mathbb{R}^n$. Then prove that $f \circ g$ is integrable on E.*

Proof. Let $E_g = \{\mathbf{x} \in E | g \text{ is not continuous at } \mathbf{x}\}$. Since f is continuous, $f \circ g$ is continuous at every point where g is continuous. Hence, if $E_{f \circ g} = \{\mathbf{x} \in E | f \circ g \text{ is not continuous at } \mathbf{x}\}$ then $E_{f \circ g} \subseteq E_g$. Since g is integrable, we know that $\lambda(E_g) = 0$, which means that $\lambda(E_{f \circ g}) = 0$ since a subset of a measure zero set has measure zero. Hence, by the Lebesgue Characterization of Riemann integrability we know that $f \circ g$ is integrable. □

Solution to Exercise 12.5. *Give an example of an integrable function f on a rectangle R in $\mathbb{R}^n$ so that the set of discontinuities of f is a dense subset of R.*

Proof. Let $f(\mathbf{x}) = 0$ if $\mathbf{x} \notin \mathbb{Q}^n \setminus \{\mathbf{0}\}$. If $\mathbf{x} \in \mathbb{Q}^n \setminus \{\mathbf{0}\}$ then let $f(\mathbf{x}) = \dfrac{1}{q}$, where q is the largest denominator of a coordinate of $\mathbf{x}$ when the coordinates of $\mathbf{x}$ are written in reduced terms.

Note that for any $k \in \mathbb{N}$ there are only finitely many elements of $\mathbf{Q}^n \setminus \{\mathbf{0}\}$ whose value if $\dfrac{1}{k}$ or larger. In particular, the n-tuples whose coordinates are all $\dfrac{i}{k}$ for $1 \leq i \leq k$ are the only points whose image under f could be $\dfrac{1}{k}$ or larger. Let $\mathbf{x}$ be a point with at least one irrational coordinate. Let $\epsilon > 0$. Choose $k \in \mathbb{N}$ so that $\dfrac{1}{k} < \epsilon$. Choose $0 < \delta$ so that $B_\delta(\mathbf{x})$ does not contain any $\mathbf{x}$ so that $f(\mathbf{x}) > \dfrac{1}{k}$. Then for all $\mathbf{y} \in B_\delta(\mathbf{x})$, we have $|f(\mathbf{x}) - f(\mathbf{y})| = f(\mathbf{y}) - 0 \leq \dfrac{1}{k} < \epsilon$. Thus, f is continuous at $\mathbf{x}$. Hence, the discontinuities of f are the points of $\mathbf{Q}^n \setminus \{\mathbf{0}\}$, which are countable, and thus the set of discontinuities of f has Lebesgue measure zero and f is integrable.

□

Solution to Exercise 12.6. *Show that for any natural number n it is possible to find $f : R \to \mathbb{R}$ which is integrable on a rectangle R in $\mathbb{R}^n$ and a function $g : I \to \mathbb{R}$ which is integrable on the interval I so that $f(R) \subseteq I$ and $g \circ f$ is not integrable on E.*

Proof. Let $R = \prod_{i=1}^{n}[0,1]$. Let $g(x) = 1$ if $x \neq 0$ and let $g(0) = 0$. Since g has only one discontinuity, g is integrable on $[0,1]$. Let $f(\mathbf{x}) = 0$ if $\mathbf{x} \notin \mathbb{Q}^n \setminus \{\mathbf{0}\}$. If $\mathbf{x} \in \mathbb{Q}^n \setminus \{\mathbf{0}\}$ then let $f(\mathbf{x}) = \frac{1}{q}$, where q is the largest denominator of a coordinate of $\mathbf{x}$ when the coordinates of $\mathbf{x}$ are written in reduced terms. It was shown in the preceding exercise that this function is integrable.

The composition $g \circ f$ is zero at every point of R with at least one rational coordinate and $g \circ f(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{Q}^n \setminus \{\mathbf{0}\}$, which is not integrable since it is discontinuous at every because every open ball contains points where $g \circ f$ takes on the values one and zero so for every point $\mathbf{x}$ it is false that there is a $\gamma > 0$ so that if $|\mathbf{x} - \mathbf{y}| < \gamma$ then $|g \circ f(\mathbf{x}) - g \circ f(\mathbf{y})| < \epsilon$. □

Solution to Exercise 12.7. *Use Change of Variables to integrate the function* $f(x,y) = x^2$ *over*

(a) The region E *enclosed by the ellipse* $\frac{x^2}{4} + \frac{y^2}{9} = 1$ *and*

(b) The region E *bounded by the lines* $y - x = 0$, $y - x = 4$, $3x + y = 0$, $3x + y = 2$.

Solution. (a) We use the transformation $x = 2u$ and $y = 3v$ so that $u^2 + v^2 \leq 1$ corresponds to the points inside the ellipse $\frac{x^2}{4} + \frac{y^2}{9} = 1$. The Jacobian of the transformation is $\begin{vmatrix} 2 & 0 \\ 0 & 3 \end{vmatrix} = 6$. Hence, the integral $\int\int_E x^2 dA = \int_D 6(2u)^2 dA$, where D is the unit disk. Converting to polar coordinates this becomes $24 \int_0^1 \int_0^{2\pi} r^3 \cos^2(\theta) d\theta dr = 24\ (\frac{r^4}{4})\Big|_0^1 (\pi) = 6\pi$.

(b) To evaluate $\int\int_E x^2 dA$ we set $u = y - x$ and $v = 3x + y$. Solving for x and y gives $x = \frac{v-u}{4}$, $y = \frac{3u+v}{4}$. This gives a Jacobian of $\begin{vmatrix} \frac{-1}{4} & \frac{1}{4} \\ \frac{3}{4} & \frac{1}{4} \end{vmatrix} = \frac{-1}{4}$. This change of variables changes the integral to $\frac{1}{16}\int_0^4 \int_0^2 \frac{1}{4}(v^2 - 2uv + u^2) dv du = \frac{1}{64}\int_0^4 \frac{v^3}{3} - uv^2 + u^2 v\Big|_0^2 du = \frac{1}{64}\int_0^4 \frac{8}{3} - 4u + 2u^2 du = \frac{1}{64}(\frac{8}{3}u - 2u^2 + \frac{2u^3}{3})\Big|_0^4 = \frac{1}{6} - \frac{1}{2} + \frac{2}{3} = \frac{1}{3}$. □

Solution to Exercise 12.8. *Let* E *be a Jordan region in* $\mathbb{R}^n$. *Prove or disprove:* E° *is also a Jordan region.*

Proof. Let $\mathbf{x} \in \partial(E^\circ)$ and let $\epsilon > 0$. Then $B_\epsilon(\mathbf{x})$ contains a point $\mathbf{y}$ of E° and a point $\mathbf{z} \notin E^\circ$. We can then find $\delta > 0$ so that $B_\delta(\mathbf{z}) \subset B_\epsilon(\mathbf{x})$ since $B_\epsilon(\mathbf{x})$ is open. Since $\mathbf{z} \notin E^\circ$ there is a point $\mathbf{w}$ of $B_\delta(\mathbf{z})$ which is not contained in E, which means that $B_\epsilon(\mathbf{x})$ contains a point which is not contained in E. Thus, $\mathbf{x} \in \partial(E)$ so $\partial(E^\circ) \subseteq \partial(E)$. Since $Vol(\partial(E)) = 0$, $Vol(\partial(E^\circ)) = 0$, which means that E° is a Jordan region.

□

Solution to Exercise 12.9. *Prove that any bounded subset E of $\mathbb{R}^n$ which has only finitely many limit points has volume zero.*

Proof. Let F be the set of limit points of E. Let $\epsilon > 0$. Since F is finite we know that F has volume zero, and that we can find cubes $C_1, C_2, ..., C_m$ whose interiors cover F so that $\sum_{i=1}^{m} |C_i| < \frac{\epsilon}{2}$ by Theorem 12.27. Then $W = E \setminus \bigcup_{i=1}^{m} C_i$ has no limit points because any limit point of W would be a limit point of E and therefore a point of F, but since every point of F is contained C_i° for some i we know that no point of F is a limit point of E. By the Bolzano-Weierstrass theorem, any bounded infinite set must have a limit point, which means that W is finite. Thus, we can find a collection of cubes $K_1, K_2, ..., K_t$ covering W so that $\sum_{i=1}^{t} |K_i| < \frac{\epsilon}{2}$. Hence, $\{C_i\}_{i=1}^m \cup \{K_i\}_{i=1}^t$ is a cover of E by cubes with the sum of the volumes of the cubes in the cover less than ϵ. So, $Vol(E) = 0$.

□

Solution to Exercise 12.10. *Prove that every open ball (and every closed ball) in $\mathbb{R}^n$ is a Jordan region.*

Proof. We begin by proving inductively that any ball about the origin in $\mathbb{R}^n$ is a Jordan region. First, the boundaries of open and closed balls are the same, so we need only verify that a closed ball is a Jordan region, from which it will follow that an open ball is a Jordan region.

If $n = 1$ the boundary of $[-r, r]$ is $\{-r, r\}$ which is a set containing two points and has measure zero, so $[-r, r]$ is a Jordan region. Assume that $\overline{B_r(\mathbf{0})} = D$ is a Jordan region in $\mathbb{R}^k$. Then the boundary of $\overline{B_r(\mathbf{0})}$ in $\mathbb{R}^{k+1}$ is the union of the graphs $G_1^{k+1} = \{(\mathbf{x}, \sqrt{r^2 - |\mathbf{x}|^2}) | \mathbf{x} \in D\}$ and $G_2^{k+1} = \{(\mathbf{x}, -\sqrt{r^2 - |\mathbf{x}|^2}) | \mathbf{x} \in D\}$, both of which are graphs of continuous functions over Jordan regions, so the boundary of $\overline{B_r(\mathbf{0})}$ has volume zero and $\overline{B_r(\mathbf{0})}$ is a Jordan region by Theorem 12.50.

Since we have proven that the translation of a Jordan region is a Jordan region, and every closed ball is a translation of a ball centered at the origin, all balls (open and closed) are Jordan regions in $\mathbb{R}^n$ for each $n \in \mathbb{N}$ by induction. □

Solution to Exercise 12.11. *Show that the countable union of Jordan regions need not be a Jordan region, even if the union is bounded.*

Proof. A single point is a Jordan region, so the points of $S = \mathbb{Q}^n \cap \overline{B_1(\mathbf{0})}$ is a countable union of Jordan regions which is not a Jordan region (since $\partial(S) = \overline{B_r(\mathbf{0})}$, which does not have outer volume zero). □

Solution to Exercise 12.12. *Let E be a Jordan region in $\mathbb{R}^n$. Define the dilation of E by a factor of $k > 0$ to be $kE = \{k\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x} \in E\}$. Prove that kE is a Jordan region and $Vol(kE) = k^n Vol(E)$.*

Proof. This follows immediately from Theorem 12.65, since kE is just AE, where A is the diagonal matrix whose diagonal entries are all k, and $\det(A) = k^n$.

□

Solution to Exercise 12.13. *Prove the Mean Value Theorem for Multiple Integrals. Let $f, g : E \to \mathbb{R}$ be integrable with $g(\boldsymbol{x}) \geq 0$ on the Jordan region E in $\mathbb{R}^n$. Then there is a number $c \in [\inf f(E), \sup f(E)]$ so that $c\int_E g = \int_E fg$. If f is continuous and E is connected then there is a value $\boldsymbol{w} \in E$ so that $f(\boldsymbol{w}) = c$. In particular, there is a number $c \in [\inf f(E), \sup f(E)]$ so that $cVol(E) = \int_E f$.*

Proof. Since $g(\mathbf{x}) \geq 0$ we know that $\inf f(E)g(\mathbf{x}) \leq f(\mathbf{x})g(\mathbf{x}) \leq \sup f(E)g(\mathbf{x})$. Thus, $\inf f(E)\int_E g \leq \int_E fg \leq \sup f(E)\int_E g$ by Theorem 6.4. If $\int_E g = 0$ then for any number c it is true that $cVol(E) = \int_E f = 0$. Otherwise, $\inf f(E) \leq \frac{\int_E fg}{\int_E g} \leq \sup f(E)$. Thus, if $c = \frac{\int_E fg}{\int_E g}$ then $c\int_E g = \int_E fg$. If E is connected and f is connected then by the Intermediate Value Theorem for $\mathbb{R}^n$ there is some $\mathbf{w} \in E$ so that $f(\mathbf{w}) = c$. In the case where $g(x) = 1$ we know that $\int_E g = Vol(E)$, so $cVol(E) = \int_E f$.

□

Solution to Exercise 12.14. *Let $f : E \to \mathbb{R}$ be integrable on the Jordan region E in $\mathbb{R}^n$, where f is continuous at the point $\boldsymbol{z} \in E^\circ$. Prove that $\lim_{r\to 0^+} \frac{1}{Vol(B_r(\boldsymbol{z}))}\int_{B_r(\boldsymbol{z})} f = f(\boldsymbol{z})$.*

Proof. Let $\epsilon > 0$. Choose $\delta > 0$ so that if $|\mathbf{z} - \mathbf{x}| < \delta$ then $|f(\mathbf{x}) - f(\mathbf{z})| < \epsilon$.

Then $f(\mathbf{z}) - \epsilon \leq f(\mathbf{x}) \leq f(\mathbf{z}) + \epsilon$ for all $\mathbf{x} \in B_\delta(\mathbf{z})$, so $(f(\mathbf{z}) - \epsilon)(Vol(B_\delta(\mathbf{z})) \leq \int_{B_\delta(\mathbf{z})} f \leq (f(\mathbf{z}) + \epsilon)(Vol(B_\delta(\mathbf{z})))$ and therefore if $0 < r < \delta$ then $f(\mathbf{z}) - \epsilon \leq \frac{1}{Vol(B_r(\mathbf{z}))}\int_{B_r(\mathbf{z})} f \leq f(\mathbf{z}) + \epsilon$. Hence, $\lim_{r\to 0^+} \frac{1}{Vol(B_r(\mathbf{z}))}\int_{B_r(\mathbf{z})} f = f(\mathbf{z})$. □

Solution to Exercise 12.15. *Let $f, g : E \to \mathbb{R}$ be integrable on the Jordan region E in $\mathbb{R}^m$. Show that if $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for every point $\boldsymbol{x} \in \mathbb{Q}^m \cap E$ then $\int_E f = \int_E g$.*

Proof. Choose a sequence of grids $\{G_n\}$ on a rectangle R containing E so that $\{|G_n|\} \to 0$. By Theorem 12.30, if we choose any markings T_n of G_n then $\{S_{T_n}(f, G_n)\} \to \displaystyle\int_E f$ and $\{S_{T_n}(g, G_n)\} \to \displaystyle\int_E g$. However, since $\mathbb{Q}^m$ is dense in every rectangle in each grid G_n, we can choose the markings T_n to consist of only points of $\mathbb{Q}^m$. Then $S_{T_n}(f, G_n) = S_{T_n}(g, G_n)$ for every $n \in \mathbb{N}$. Hence, $\displaystyle\int_E g = \int_E f$.

□

Chapter 13

Vector Fields, Curves and Surfaces

In this chapter, we discuss curves, surfaces, vector fields and integrals along curves and surfaces. We will begin with a discussion of curves.

Curves

Definition 98

For any $k \in \mathbb{N}$ or for $k = \infty$, if there is an open interval I containing D and a C^k function $\mathbf{r}^* : I \to \mathbb{R}^n$ so that $\mathbf{r}^*$ restricted to D is $\mathbf{r}$, then we say that $\mathbf{r}$ is C^k. If $\mathbf{r}$ is C^1 and $\mathbf{r}'(t) \neq 0$ for each $t \in D$, and one to one on D°, then we refer to $\mathbf{r}$ as a *smooth* curve. We also refer to the trace of C as a smooth curve if there is any parametrization whose trace is C which is a smooth curve. If $D = [a, b]$ then we refer to $\mathbf{r}(a)$ and $\mathbf{r}(b)$ as the *end points* (the initial and terminal or starting and ending points respectively) of C. If $\mathbf{r}$ is a smooth curve which is one to one then we refer to $\mathbf{r}$ as a *simple smooth curve*. If $\mathbf{r}$ is a smooth path with $D = [a, b]$ so that $\mathbf{r}(a) = \mathbf{r}(b)$ then we will refer to C as a *smooth closed curve*. If C is a smooth closed curve so that $\mathbf{r}'(a) = \mathbf{r}'(b)$ then we refer to C as a *smooth simple closed curve.* A *regular* curve is a C^∞ curve $\mathbf{r}$ so that $\mathbf{r}'$ is never zero (but a regular curve need not be one to one). The trace of a curve is an *arc* if $\mathbf{r}$ has domain $[a, b]$ and $\mathbf{r}$ is one to one. The trace is a *simple closed curve* if $\mathbf{r}$ has domain $[a, b]$, $\mathbf{r}(a) = \mathbf{r}(b)$ and the function is otherwise one to one (meaning that if $x \notin \{a, b\}$ and $x \neq y$ then $f(x) \neq f(y)$).

It should be noted that what we have defined to be a curve is usually not all that useful except for determining things like path connectedness. It is known that there are space filling curves for instances, which are continuous images of closed intervals that are two higher (they fill a square an n-rectangle and do not look like curves at all). However, we sometimes want to talk about such things too, so we are using the term ”curve” to be this most general sort of curve, in keeping with the notion of a path in topology. Mostly, we care about smooth curves. Without the $\mathbf{r}'(t) \neq 0$ requirement there may be no tangent lines to a curve and the graph may have cusps. Without the one to one requirement, notions of arc length become confusing (they end up referring to distance traveled along a curve, possibly moving back and forth, rather than the length of the curve itself).

Definition 99

We say a set C in $\mathbb{R}^n$ is an *arc* if $C = f([a, b])$ for some closed interval $[a, b]$ and some continuous one to one function $f : [a, b] \to \mathbb{R}^n$. A *simple closed curve* is the one to one continuous image of a circle.

Since an arc and a simple closed curve do not have to be differentiable, they are usually not the objects were are interested in if our goal is to explore calculus notions on an object. However, there are results which are about arcs or simple closed curves that are relevant to our discussions.

Definition 100

Let C be the one to one curve $\mathbf{r} : [a_0, a_m] \to \mathbb{R}^n$. If there are simple smooth curves $C_1 = \mathbf{r}_1([a_0, a_1]), C_2 = \mathbf{r}_2([a_1, a_2]), ..., C_m = \mathbf{r}_m([a_{m-1}, a_m])$ so that $\mathbf{r}(x) = \mathbf{r}_i(x)$ whenever $x \in [a_{i-1}, a_i]$ then we say that C is a *piecewise- smooth simple curve* and we will refer to the set $\{C_1, ..., C_m\}$ as a *decomposition* for C, and each C_i as a *component curve* of C. If we amend the definition of piecewise -smooth simple curve so that $r_1(a_1) = r_m(b_m)$ (and $\mathbf{r}$ is otherwise one to one, meaning that if $x \notin \{a_0, a_m\}$ and $y \in [a_0, a_m]$ then $\mathbf{r}(x) \neq \mathbf{r}(y)$) then we say C is a *piecewise-smooth closed curve*.

The reader should be aware that there are some differences in the specifics what what some of the terms above mean from one text to another. For example, in some books a smooth curve may only have to continuously differentiable, or a smooth curve may have to be regular. These definitions aren't consistent.

Example 13.1. *(a) Explain why the parametrization* $\mathbf{r}(t) =< a + R\ cos(t), b + R \sin(t) >$, $0 \leq t \leq 2\pi$ *is a parametrization for a circle of radius* R *centered at* (a, b)*, traversed once counterclockwise.*

(b) What curve would be traversed by $\mathbf{r}(t) =< a\cos(t), b\cos(t) >$, for $0 \leq t \leq 2\pi$?

(c) What curve would be traversed by $\mathbf{r}(t) =< \cos(t), \cos(t), t >$, for $t \in \mathbb{R}$?

Solution. (a) We observe that $(x - a)^2 + (y - b)^2 = R^2\cos^2(t) + R^2\sin^2(t) = R^2$, which we know is the equation for a circle of radius R centered at (a, b). For each point $\mathbf{p}$ on the circle, there is a unique angle $\theta \in [0, 2\pi)$ so that a radius R line segment from center (a, b) at angle θ would end at $\mathbf{p}$, which means that $\mathbf{p} =< a + R\ cos(\theta), b + R\sin(\theta) >$. In the interval $[0, 2\pi]$ the only repeated point would be at 0 and 2π, so the circle is traversed only once. The circle is traversed counterclockwise since larger angle values measured from the positive x direction correspond to points further along the curve in the counterclockwise direction.

(b) In the case of $\mathbf{r}(t) =< a\cos(t), b\cos(t) >$, we note that $\dfrac{x^2}{a^2}+\dfrac{y^2}{b^2} = \cos^2(t)+\sin^2(t) = 1$, so the curve traced out is an ellipse.

(c) Since $x^2 + y^2 = 1$ just as it was in (a), the points of the parametrized curve lie on the cylinder with radius one about the z-axis. The third component increases at a constant rate, which gives a trace which is a helix twisting upwards.

□

It is worth noting that there are multiple parametrizations for the same curve. For instance, the circle in example (a) can be traversed by $\mathbf{r}(t) =< a+R\cos(2t), b+R\sin(2t) >$, $0 \leq t \leq \pi$, which is a parametrization tracing out the same curve counterclockwise at twice the speed of the former parametrization. We would like to formalize notions of speed and acceleration of a particle moving along a parametrized path, motivating the following definitions.

Let $\mathbf{r}(t) =< x(t), y(t), z(t) >$ be a parametrized curve defined on an interval I. We should observe that, for each point $t_0 \in I$, $\lim\limits_{t\to t_0} \mathbf{r(t)} =< x_0, y_0, z_0 >$ if and only if $\lim\limits_{t\to t_0} x(t) = x_0$, $\lim\limits_{t\to t_0} y(t) = y_0$ and $\lim\limits_{t\to t_0} z(t) = z_0$. Likewise, for a two component parametrized curve $\mathbf{r}(t) =< x(t), y(t) >$ defined on an interval I, for each point $t_0 \in I$ it is the case that $\lim\limits_{t\to t_0} \mathbf{r(t)} =< x_0, y_0 >$ if $\lim\limits_{t\to t_0} x(t) = x_0$ and $\lim\limits_{t\to t_0} y(t) = y_0$.

Example 13.2. *Find* $\lim\limits_{t\to 0} < \dfrac{\sin(3t)}{t}, \dfrac{e^t - 1 - t}{t}, \dfrac{\cos(t) - 1}{t^2} >$.

Solution. We can just take the limit of each coordinate and use L'Hospital's Rule. For the first coordinate we get $\lim\limits_{t\to 0} \dfrac{\sin(3t)}{t} = \lim\limits_{t\to 0} \dfrac{3\cos(3t)}{1} = 3$. In the second coordinate, $\lim\limits_{t\to 0} \dfrac{e^t - 1 - t}{t} = \lim\limits_{t\to 0} \dfrac{e^t - 1}{1} = 0$ and in the third coordinate $\lim\limits_{t\to 0} \dfrac{\cos(t) - 1}{t^2} = \lim\limits_{t\to 0} \dfrac{-\sin(t)}{2t} = \lim\limits_{t\to 0} \dfrac{-\cos(t)}{2} = -\dfrac{1}{2}$. Thus, the limit is $< 3, 0, -\dfrac{1}{2} >$.

□

Example 13.3. *Find the derivative of* $\boldsymbol{r}(t) =< t, e^t, t^3 + 1 >$ *at* $(0, 1, 1)$.

Solution. Setting the first coordinates equal to each other we see $t = 0$. $\mathbf{r}'(t) =< 1, e^t, 3t^2 >$. Setting $t = 0$ we see $\mathbf{r}'(0) =< 1, 1, 0 >$.

□

It is often helpful to know the length of a curve. This can help us to determine how far an object moves on when it traverses a curve over a given time interval.

Definition 101

Let $\mathbf{r}(t)$ on $a \leq t \leq b$ be a regular curve. Let $\{P_n\}$ be the standard sequence of partitions of $[a,b]$ into n evenly spaced subdivisions. We define $L = \lim\limits_{n\to\infty} \sum\limits_{i=1}^{n} |\mathbf{r}(t_i) - \mathbf{r}(t_{i-1})|$ to be the *arc length* of the trace of this regular curve. If C has a length which is finite then we say that C is *rectifiable.*

Theorem 13.1. *Let $\boldsymbol{r} : [a,b] \to \mathbb{R}^3$ be a smooth curve with trace C. Then C is rectifiable with arc length $L = \int_a^b |\boldsymbol{r}'(t)|dt$.*

In particular:

If $\boldsymbol{r}(t) =< x(t), y(t) >$ over $a \leq t \leq b$ then the arc length of C is given by the formula $L = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}dt$.

If $\boldsymbol{r}(t) =< x(t), y(t), z(t) >$ over $a \leq t \leq b$ then the arc length of C is given by the formula $L = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2 + (z'(t))^2}dt$.

If $y = f(x)$ over $a \leq x \leq b$ then $L = \int_a^b \sqrt{1 + (f'(x))^2}dx$.

Furthermore, there is a $\delta > 0$ so that if $P = \{t_0, t_1, ..., t_n\}$ is a partition of $[a,b]$ with $|P| < \delta$ then $| \int_a^b |\boldsymbol{r}'(t)|dt - \sum\limits_{i=1}^{n} |\boldsymbol{r}(t_i) - \boldsymbol{r}(t_{i-1})|| < \epsilon$.

Proof. We will prove the theorem for a curve $\mathbf{r}(t) =< x(t), y(t), z(t) >$, from which the result will follow for a two coordinate curve (by setting $z(t) = 0$).

First, note that $L = \int_a^b |\mathbf{r}'(t)|dt$ exists since $|\mathbf{r}'(t)|$ is continuous.

Let $P_n = \{t_0, t_1, ..., t_n\}$ be the standard nth partition of $[a,b]$ into equal length subintervals with each $t_i - t_{i-1} = \Delta t$. In each subinterval $[t_{i-1}, t_i]$, by the Mean Value Theorem, we can pick $t_i^(x), t_i^{(y)}, t_i^{(z)}$ so that $x'(t_i^{(x)})\Delta t = x(t_i) - x(t_{i-1})$, $y'(t_i^{(y)})\Delta t = y(t_i) - y(t_{i-1})$, $z'(t_i^{(z)})\Delta t = z(t_i) - z(t_{i-1})$. The distance from $\mathbf{r}(t_{i-1})$ to $\mathbf{r}(t_i)$ is

$\sqrt{(x(t_i) - x(t_{i-1}))^2 + (y(t_i) - y(t_{i-1}))^2 + (z(t_i) - z(t_{i-1}))^2}$

$= \sqrt{(x'(t_i^{(x)}))^2(\Delta t)^2 + (y'(t_i^{(y)}))^2(\Delta t)^2 + (z'(t_i^{(z)}))^2(\Delta t)^2}$

$= \sqrt{(x'(t_i^{(x)}))^2 + (y'(t_i^{(y)}))^2 + (z'(t_i^{(z)}))^2}\Delta t.$

Let $G(u,v,w) = \sqrt{x'(u)^2 + y'(v)^2 + z'(w)^2}$. Note that G is continuous since each of x', y', z' are continuous. Thus, on the closed and bounded set $[a,b] \times [a,b] \times [a,b]$ we know that G is uniformly continuous by Theorem 10.35. Choose a $\delta > 0$ so that if $|\mathbf{x} - \mathbf{y}| < \delta$ then $|G(\mathbf{x}) - G(\mathbf{y})| < \frac{\epsilon}{b-a}$. If n is large enough that $\Delta t < \frac{\delta}{\sqrt{3}}$ it follows that each $|(t_i, t_i, t_i) - (t_i^{(x)}, t_i^{(y)}, t_i^{(z)})| < \delta$ which means that

$|\sum\limits_{i=1}^{n} \sqrt{(x'(t_i^{(x)}))^2 + (y'(t_i^{(y)}))^2 + (z'(t_i^{(z)}))^2}\Delta t$ -

$\sum_{i=1}^{n} \sqrt{(x'(t_i))^2 + (y'(t_i))^2 + (z'(t_i))^2}\Delta t| = |\sum_{i=1}^{n}(G(t_i,t_i,t_i) - G(t_i^{(x)}, t_i^{(y)}, t_i^{(z)})\Delta t| \le \sum_{i=1}^{n} |G(t_i,t_i,t_i) -$ $G(t_i^{(x)}, t_i^{(y)}, t_i^{(z)})|\Delta t \le (b-a)\frac{\epsilon}{(b-a)} = \epsilon.$

We know that $\lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(x'(t_i))^2 + (y'(t_i))^2 + (z'(t_i))^2}\Delta t =$ $\int_a^b |\mathbf{r}'(t)|dt = L$ and $\lim_{n\to\infty} \sum_{i=1}^{n} \sqrt{(x'(t_i))^2 + (y'(t_i))^2 + (z'(t_i))^2} -$ $\sum_{i=1}^{n} \sqrt{(x'(t_i^{(x)}))^2 + (y'(t_i^{(y)}))^2 + (z'(t_i^{(z)}))^2}\Delta t = 0.$ Hence, $\lim_{n\to\infty} \sum_{i=1}^{n} |\mathbf{r}(t_i) - \mathbf{r}(t_{i-1})|$ $= \lim_{n\to\infty} \sqrt{(x'(t_i^{(x)}))^2 + (y'(t_i^{(y)}))^2 + (z'(t_i^{(z)}))^2}\Delta t = L.$ Thus, C is rectifiable with arc length $L = \int_a^b |\mathbf{r}'(t)|dt.$

In particular if we parametrize $y = f(x)$ on the interval $[a, b]$ with the parametrization $\mathbf{r}(t) =< t, f(t) >$ for $a \le t \le b$ then we have $L = \int_a^b \sqrt{1^2 + (f'(t))^2}dt = \int_a^b \sqrt{1 + (f'(x))^2}dx.$

To verify the last part of the theorem, we note that none of the steps in the argument above depended on the partition points being equally spaced, only on the mesh of the partition being sufficiently small.

□

We know that $s = \int_a^t |\mathbf{r}'(t)|dt$ exists for any smooth curve s, which is the *arc length of the curve C parametrized by $\mathbf{r}(t)$ measured from the point $\mathbf{r}(a)$*. Since $|\mathbf{r}'(t)|$ is positive and continuous, we know that $s(t)$ is an increasing function of t. Thus, $s(t)$ is invertible, with inverse $t(s)$. We say that the function $\mathbf{r}(t(s))$ or just $\mathbf{r}(s)$ (where it is understood that $\mathbf{r}(s)$ refers to $\mathbf{r} \circ t(s)$) is a *parametrization of C with respect to arc length*. If we fix the value of a then this parametrization is unique.

Definition 102

Let $\mathbf{r}(t)$ be parametrized curve defined on an open interval I. The *velocity* $\mathbf{v}(t) = \mathbf{r}'(t)$ for this curve, or the velocity of an object whose position is $\mathbf{r}(t)$, and the *acceleration* of an object whose position is $\mathbf{r}(t)$ is $\mathbf{r}''(t) = \mathbf{a}(t)$ (assuming these derivatives exist). We say that $|\mathbf{v}(t)| = v(t)$ is the *speed* of a particle whose position at time t is $\mathbf{r}(t)$. Thus, $v(t) = |\mathbf{r}'(t)|$.

The *integral* of the vector valued function $\mathbf{r}(t) =< x_1(t), x_2(t), ...x_n(t) >$ over $[a, b]$ is $< \int_a^b x_1(t)dt, \int_a^b x_2(t)dt, ..., \int_a^b x_n(t)dt >$.

If the curve C parametrized by $\mathbf{r}(t)$ is smooth then $s(t) = \int_a^t |\mathbf{r}'(t)|dt$ is the *arc length function* measured from the point $\mathbf{r}(a) = \mathbf{p}$. If $t(s)$ is the inverse of $s(t)$ then $\mathbf{r}(s)$ denotes $\mathbf{r}(t(s))$ is the *parametrization of C with respect to arc length* measured from the point **p**.

Example 13.4. *Find the integral of* $\mathbf{r}(t) =< e^t, 2t, 4 >$ *over* $[0, 2]$.

Solution. The integral is $< \int_0^2 e^t dt, \int_0^2 2t dt, \int_0^2 4 dt > = < e^t, t^2, 4t >|_0^2 =< e^2 - 1, 4, 8 >$. □

Using parametric equations it is possible to find formulas for area under a curve and for the rate of change of one variable with respect to another.

Theorem 13.2. *Let* $\mathbf{r}(t) =< x(t), y(t) >$ *be a* C^1 *parametrized curve defined on an open interval* I*. If* $x'(t_0) \neq 0$ *for some* $t_0 \in I$ *then for some* $\delta > 0$*, if we restrict* $\mathbf{r}(t)$ *to the domain* $(t_0 - \delta, t_0 + \delta)$ *then* y *is a differentiable function of* x *on that curve and* $\frac{dy}{dx}$ *exists at* $x(t_0)$ *and is equal to* $\frac{y'(t_0)}{x'(t_0)}$*. Furthermore, the second derivative* $\frac{d^2y}{dx^2} = \frac{(\frac{y'(t)}{x'(t)})'(t_0)}{x'(t_0)}$

Proof. First, we note that we can choose a $\delta > 0$ so that if $|t - t_0| < \delta$ then $|x'(t) - x'(t_0)| < \frac{1}{2}|x'(t_0)|$ which means that either $x'(t) > \frac{1}{2}x'(t_0) > 0$ or $x'(t) < \frac{1}{2}x'(t_0) < 0$ on $(t_0 - \delta, t_0 + \delta)$. Thus, $x(t)$ is strictly monotone on $(t_0 - \delta, t_0 + \delta)$. Hence, for every $t \in (t_0 - \delta, t_0 + \delta)$, it follows that there is only one y value $y(t)$ so that $(x(t), y(t)) \in \mathbf{r}((t_0 - \delta, t_0 + \delta))$. This means that y is a function of x. Furthermore, since $x(t)$ is continuously differentiable, $x(t)$ has a continuously differentiable inverse function $t(x)$ so that $t'(x) = \frac{1}{x'(t)}$ by the Inverse Function Theorem (where t is the point that maps to x under $x(t)$). Using the chain rule and the Inverse Function Theorem we see that $\frac{dy}{dx}$ is the derivative of $y(t(x))$ with respect to x, which is $y'(t(x))t'(x) = y'(t)\frac{1}{x'(t)}$.

To prove the last part of the theorem, we note that since $\frac{y'(t)}{x'(t)}$ is $\frac{dy}{dx}(t)$ on $(t_0 - \delta, t_0 + \delta)$, by the first part of the argument (substituting $\frac{dy}{dx}(t)$ for $y(t)$), it follows that the rate of change of $\frac{dy}{dx}$ with respect to x is $\frac{(\frac{y'(t)}{x'(t)})'(t_0)}{x'(t_0)}$ at $x(t_0)$. □

It is important to notice in the preceding discussion that the value of $\frac{dy}{dx}$ obtained only exists locally at t_0, meaning that it is only for the function restricted to the $(t - \delta, t_0 + \delta)$ interval (or any interval on which $x(t)$ is strictly monotone which contains t_0). If there is only one t value at which $x(t) = x(t_0)$ or if $\frac{dy}{dx}$ is the same for all t values so that $x(t) = x(t_0)$ then there is such a thing as the tangent line slope at $x(t_0)$ for $\mathbf{r}(t)$, and

$\frac{dy}{dx}(x(t_0))$ is the tangent line slope. On the other hand, if the graph of $\mathbf{r}(t)$ passes through the same point at two different times t_1, t_2 and the slopes $\frac{dy}{dx}$ exist locally at both values of t but $\frac{dy}{dx}(x(t_1)) \neq \frac{dy}{dx}(x(t_2))$ (the values of $\frac{dy}{dx}$ for portions of the curve at t values near t_1 and t_2 respectively) then the graph of $\mathbf{r}(t)$ does not have a tangent line at $(x(t_1), y(t_1))$. In some texts, it is said that the graph has two tangent lines at the point but we will use the convention that seems more common in later advanced calculus courses, and say that the curve restricted to two subintervals of its domain has two different tangent lines but the entire graph has none. Another way of saying this is:

Definition 103

If the parametrized curve $\mathbf{r} : (a, b) \to \mathbb{R}^2$ has a trace which, if intersected with some $B_\epsilon(\mathbf{r}(t_0))$, is also the graph of a function $y = f_1(x)$ or $x = f_2(y)$ for differentiable functions f_1 or f_2 then the trace of the curve is *locally a differentiable function graph near* $\mathbf{r}(t_0)$, and a *tangent line* to the trace of $\mathbf{r}((a, b))$ exists. In the case where $x'(t_0) = 0$ the tangent line is vertical (but still exists even though $\frac{dy}{dx}$ does not).

By simply switching variable labels, we can likewise find $\frac{dx}{dy}(y(t_0)) = \frac{x'(t_0)}{y'(t_0)}$ for the parametrization restricted to an interval of t values about t_0 on which $y(t)$ is monotone, so we can identify vertical tangent lines when $x'(t) = 0$, for instance, just as horizontal tangent lines occur when $y'(t) = 0$.

Example 13.5. *Let* $\mathbf{r}(t) =< t^3 + 2t, e^{2t} >$*. Find the tangent line to this curve, and also find* $\frac{d^2y}{dx^2}$ *at* $(0, 1)$*.*

Solution. This point occurs when $t = 0$. Note that $x'(0) = 3(0^2 + 2 > 1$ so near $t = 0$ it is true that y is a function of t. In fact, $x'(t) > 0$ for all t, which means that there is no other value t so that $x(t) = 0$, so there is a tangent line to the curve (whether the domain is restricted or not) at each point. The derivative is $\frac{dy}{dx} = \frac{2e^{2t}}{3t^2 + 2} = \frac{2}{2} = 1$ when $t = 0$. The second derivative is $\frac{d^2y}{dx^2} = \frac{\frac{(3t^2+2)(4e^{2t}) - 2e^{2t}(6t)}{(3t^2+2)^2}}{3t^2 + 2}$. At $t = 0$ this is also equal to 1.

Since $\frac{dy}{dx}$ is the slope of the tangent line, we can use the usual point-slope form of the line to get the tangent line, which is $y - 1 = (1)(x - 0)$.

□

We can also use parametrizations to find the area between a parametrized curve and an axis, or enclosed by a parametrized simple closed curve. We discuss the latter more when we talk about Green's Theorem.

Theorem 13.3. *Let $\mathbf{r}(t) =< x(t), y(t) >$ be a C^1 parametrized curve defined on an open interval I so that on a closed interval $[a, b] \subset I$ it is true that $x'(t) > 0$ on the interior of I. Then $\int_{x(a)}^{x(b)} y(t(x))dx = \int_a^b y(t)x'(t)dt$. If $x'(t) < 0$ on $[a, b]$ then $\int_{x(b)}^{x(a)} y(t(x))dx = -\int_a^b y(t)x'(t)dt$.*

Proof. First, note that since $y(t)x'(t)$ is continuous, it is integrable. As in the preceding theorem, we know that y is a function of x on $[a, b]$ since $x(t)$ is strictly monotone, so if $t(x)$ is the inverse function of $x(t)$ then $y(t(x))$ is a function of x. Using the substitution $x = x(t)$ we have that $dx = x'(t)dt$. Hence, $\int_a^b y(t)x'(t)dt = \int_a^b y(t(x(t))x'(t)dt = \int_{x(a)}^{x(b)} y(t(x))dx$. In the case of $x'(t) < 0$ we would have bounds where $x(a) > x(b)$, and switching them negates the integral. □

Example 13.6. *Find the area enclosed by the ellipse $\mathbf{r}(t) =< a\cos(t), b\sin(t) >$, $0 \le t \le 2\pi$.*

Solution. Since $x'(t) < 0$ on the interior of the interval $[0, \frac{\pi}{2}]$, and the portion of the ellipse in the first quadrant is traced out by the parametrization restricted to this subinterval. From this, it follows that the enclosed area is $-4\int_0^{\frac{\pi}{2}} b\sin(t)(-a\sin(t))dt$. Using Wallis's formula we obtain $4\int_0^{\frac{\pi}{2}} ab\sin^2(t)dt = 4ab\frac{1}{2}\frac{\pi}{2} = \pi ab$.

□

Definition 104

Let $\mathbf{r}(t)$ be a smooth curve. The *unit tangent vector* is $\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{|\mathbf{r}'(t)|}$. This is a unit vector in the direction of the curve.

The *unit normal vector* $\mathbf{N}(t) = \frac{\mathbf{T}'(t)}{|\mathbf{T}'(t)|}$. This vector is perpendicular to the direction of a curve, and the curve accelerates into the normal direction as well as accelerating along the direction of the tangent vector. All of the acceleration of $\mathbf{r}(t)$ is a sum of its acceleration into the tangent direction and its acceleration in the direction of the unit normal vector. The fact that $\mathbf{T}(t)$ and $\mathbf{N}(t)$ are perpendicular just follows taking $\mathbf{T} \cdot \mathbf{T} = 1$ and differentiating both sides to get $2\mathbf{T}' \cdot \mathbf{T} = 0$, so $\mathbf{T}$ is perpendicular to $\mathbf{T}'$.

The *unit binormal* vector is $\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t)$.

The *curvature* κ of $\mathbf{r}(t)$ is $|\frac{d\mathbf{T}}{ds}|$.

Let $\alpha(s)$ be a regular curve parametrized with respect to arc length. The *torsion* $\tau(s)$ is the number so that $\mathbf{B}'(s) = \tau(s)\mathbf{N}(s)$.

The *normal plane* to a curve $\mathbf{r}(t)$ at time $t = t_0$ is the plane perpendicular to $\mathbf{r}'(t_0)$ which passes through the point $\mathbf{r}(t_0)$. The *osculating plane* is the plane which is perpendicular to the binormal vector at time $t = t_0$ and passes through the point $\mathbf{r}(t_0)$. The *rectifying plane* is the plane through the same point which is perpendicular to the unit normal vector.

Initially, it might seem like this definition of curvature could depend on which parametrization we use, but any parametrization with respect to arc length gives the same curvature definition. In fact, we will show this definition of curvature is the same as $\dfrac{|\mathbf{T}'(t)|}{|\mathbf{r}'(t)|} = \dfrac{|\mathbf{r}'(t) \times \mathbf{r}''(t)|}{|\mathbf{r}'(t)|^3}$. Typically, we will want to use the $\dfrac{|\mathbf{r}'(t) \times \mathbf{r}''(t)|}{|\mathbf{r}'(t)|^3}$ formula to find curvature. Curvature is a measure of how sharply a space curve turns. A radius R circle has curvature $\dfrac{1}{R}$, so if the curvature of a space curve is κ at a given time, then a circle of radius $\dfrac{1}{\kappa}$ would bend as sharply as the given curve at that point along the curve.

Note that in some books the torsion we gave is the negative of the torsion rather than the torsion. That there even is such a number requires a theorem, and is shown below.

Note that the unit tangent, normal, and binormal vectors, as well as the curvature, are local properties of a parametrized curve. Thus, if a curve intersects itself later then at a given point on the trace of the curve there could be two unit tangent vectors, one at the time value the first time the curve passed through the point, and another unit tangent vector at another time. Hence, we talk about the unit tangent, normal, binormal vectors and curvature at a value t of the parameter rather than at a point on the trace of the curve itself.

Theorem 13.4. *Let $\boldsymbol{r} : I \to \mathbb{R}^3$ be a smooth curve, where $t_0 \in I$. Let $\boldsymbol{u}(s) : J \to \mathbb{R}^3$ be a parametrization of $\boldsymbol{r}$ with respect to arc length, where s is the arc length measured from some point $\boldsymbol{r}(t_0)$. Then $\boldsymbol{u}(s)$ is a regular curve and $\boldsymbol{u}'(s)$ is the unit tangent vector to $\boldsymbol{u}(s)$ at $s = s_0$.*

Proof. Since $s(t) = \displaystyle\int_{t_0}^{t} |\mathbf{r}'(t)|dt$ by the Fundamental Theorem of Calculus (first form) we note that $s'(t) = |\mathbf{r}'(t)| > 0$, which means that s is continuously differentiable and increasing, and thus one to one. It follows from the Inverse Function Theorem that $t(s)$, the inverse function, is also increasing and continuously differentiable. Since $\mathbf{u}(s) = \mathbf{r}(t(s))$, which is a composition of two C^∞ functions, it follows from the chain rule that $\mathbf{u}'(s) = \mathbf{r}'(t(s))t'(s)$ is C^∞. Also, $|\mathbf{u}'(s)| = |t'(s)||\mathbf{r}'(t(s))| > 0$, so $\mathbf{u}$ is a regular curve. Finally, by the Inverse Function Theorem we know that $t'(s) = \dfrac{1}{s'(t(s))}$. Thus, $|\mathbf{u}'(s)| = |t'(s)||\mathbf{r}'(t(s))|$ $= |\dfrac{1}{s'(t(s))}|r'(t(s))| = \dfrac{1}{|r'(t(s))|}|r'(t(s))| = 1$. This means that $\mathbf{T}(s_0) = \dfrac{\mathbf{u}'(s_0)}{|\mathbf{u}'(s_0)|} = \mathbf{u}'(s_0)$. □

Theorem 13.5. *Let $\boldsymbol{r} : I \to \mathbb{R}^3$ be a smooth curve, where $t_0 \in I$. Let $\boldsymbol{u}(s) : J \to \mathbb{R}^3$ be a parametrization of $\boldsymbol{r}$ with respect to arc length with $s(t) = \displaystyle\int_{t_0}^{t} |\boldsymbol{r}'(t)|dt$, and $t_1 \in I$.*

Then the curvature κ at that point (meaning $\kappa(t_1)$ or $\kappa(s(t_1))$ depending on whether we use parametrization $\boldsymbol{r}$ or $\boldsymbol{u}$) is $\dfrac{d\boldsymbol{T}}{ds}(s(t_1)) = \dfrac{\boldsymbol{T}'(t_1)}{|\boldsymbol{r}'(t_1)|} = \dfrac{|\boldsymbol{r}'(t) \times \boldsymbol{r}''(t)|}{|\boldsymbol{r}'(t)|^3} = |u''(s_1)|$.

Proof. First $|\dfrac{d\mathbf{T}}{ds}(s(t_1))|$ is the definition of $\kappa(t_1)$ and also the definition of $\kappa(s_1)$ with respect to the parametrizations $\mathbf{r}$ and $\mathbf{u}$ respectively. Parametrizing with respect to t we see from Theorem 13.2 that $|\dfrac{d\mathbf{T}}{ds}(s(t_1))| = |\dfrac{\frac{d\mathbf{T}}{dt}(t_1)}{\frac{ds}{dt}(t_1)}| = \dfrac{|\mathbf{T}'(t_1)|}{|\mathbf{r}'(t_1)|} = \kappa$. Note that it follows from this that the curvature is independent of parametrization choice.

Since $\mathbf{u}'(s_1) = \mathbf{T}(s_1)$ we have that $|\mathbf{u}''(s_1)| = |\dfrac{d\mathbf{T}}{ds}(s_1)| = \kappa$.

Let $v(t) = |\mathbf{r}'(t)|$. Then at any time t is is the case that $\mathbf{r}' = \mathbf{T}v$, which means that $\mathbf{r}'' = \mathbf{T}'v + \mathbf{T}v'$. Recall that $\mathbf{T}$ and $\mathbf{T}'$ are perpendicular (explained after the definition of unit normal vector earlier). Taking the cross product, and using the fact that if two vectors are parallel their cross product is the zero vector and if two vectors are perpendicular then the norm of their cross product is the product of their norms, we see that $\mathbf{r}' \times \mathbf{r}'' = v^2(\mathbf{T} \times \mathbf{T}')$, and since $\mathbf{T}' = \kappa v$, it follows that $|\mathbf{r}' \times \mathbf{r}''| = v^2|\mathbf{T}'||\mathbf{T}| = |v^3|\kappa$, so $\kappa = \dfrac{|\mathbf{r}' \times \mathbf{r}''|}{|\mathbf{r}'(t)|^3}$. □

The following relationships between unit tangent, unit normal and unit binormal vectors tell us quite a bit about the behavior of curves and are useful in differential geometry developments.

Theorem 13.6. *Frenet formulas: Let $\boldsymbol{u}(s)$ be a regular curve parametrized with respect to arc length. Then:*

(a) $\boldsymbol{T}'(s) = k(s)\boldsymbol{N}(s)$

(b) There is a number $\tau(s)$ (the torsion) so that $\boldsymbol{B}'(s) = \tau(s)\boldsymbol{N}(s)$

(c) $\boldsymbol{N}'(s) = -k(s)\boldsymbol{T}(s) - \tau(s)\boldsymbol{B}(s)$

Proof. (a) $k(s) = |u''(s)| = |\mathbf{T}'(s)|$ and $\mathbf{N}(s) = \dfrac{\mathbf{T}'(s)}{|\mathbf{T}'(s)|}$, so the result follows.

(b) We know $\mathbf{B}(s)$ is perpendicular to $\mathbf{N}(s), \mathbf{T}(s)$ (since the cross product of two vectors is perpendicular to each vector, and $\mathbf{B}(s) = \mathbf{T}(s) \times \mathbf{N}(s)$). Likewise, since $\mathbf{B}(s)$ is a unit vector we have $\mathbf{B}(s) \cdot \mathbf{B}(s) = 1$. Differentiating both sides we see that $2\mathbf{B}'(s) \cdot \mathbf{B}(s) = 0$, so $\mathbf{B}'(s)$ is perpendicular to $\mathbf{B}(s)$. Thus, $\mathbf{B}'(s)$ is in the osculating plane. Using the derivative formula for cross product we have $\mathbf{B}'(s) = (\mathbf{T}(s) \times \mathbf{N}(s))' = \mathbf{T}'(s) \times \mathbf{N}(s) + \mathbf{T}(s) \times \mathbf{N}'(s)$. Recall from (a) that $\mathbf{T}'(s)$ is parallel to $\mathbf{N}(s)$, and is therefore perpendicular to $\mathbf{B}'(s)$. Hence, $\mathbf{T}(s) \cdot \mathbf{B}'(s) = \mathbf{0}$. Thus, $\mathbf{B}'(s)$ is perpendicular to both $\mathbf{T}(s)$ and $\mathbf{B}(s)$ which means $\mathbf{B}(s)$ is parallel to $\mathbf{N}(s)$ and there is a number τ as described.

(c) By Theorem 10.4 part (g) we see that $\mathbf{N}(s) = \mathbf{B}(s) \times \mathbf{T}(s)$ and $\mathbf{B}(s) \times \mathbf{N}(s) = -\mathbf{T}(s)$. Thus, $\mathbf{N}'(s) = \mathbf{B}'(s) \times \mathbf{T}(s) + \mathbf{B}(s) \times \mathbf{T}'(s)$. Since $\mathbf{B}'(s) = \tau(s)\mathbf{N}(s)$, $\mathbf{B}'(s) \times \mathbf{T}(s) = -\tau\mathbf{B}(s)$, and since $\mathbf{T}'(s) = k(s)\mathbf{N}(s)$ it follows that the second vector in this sum is $\mathbf{B}(s) \times \mathbf{T}'(s) = -k(s)\mathbf{T}(s)$ since $\mathbf{B}(s) \times \mathbf{N}(s) = -\mathbf{T}(s)$. □

Example 13.7. *Let* $\boldsymbol{r}(t) =< 3\cos(t), 3\sin(t), 4t >$.

(a) Find $\mathbf{T}(t), \mathbf{N}(t), \mathbf{B}(t)$.

(b) Find $\kappa(t)$, the curvature at time t.

(c) Find the normal plane, osculating plane and rectifying plane to $\mathbf{r}(t)$ at the point $(-3, 0, 4\pi)$.

Solution:

(a) $\mathbf{T}(t) = \dfrac{\mathbf{r}'(t)}{|\mathbf{r}'(t)|} =< \dfrac{-3\sin(t)}{5}, \dfrac{3\cos(t)}{5}, \dfrac{4}{5} >$.

$\mathbf{N}(t) = \dfrac{\mathbf{T}'(t)}{|\mathbf{T}'(t)|} =< -\cos(t), -\sin(t), 0 >$.

$$\mathbf{B}(t) = \mathbf{T}(t) \times \mathbf{N}(t) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{-3\sin(t)}{5} & \dfrac{3\cos(t)}{5} & \dfrac{4}{5} \\ -\cos(t) & -\sin(t) & 0 \end{bmatrix} =< \frac{4\sin(t)}{5}, \frac{-4\cos(t)}{5}, \frac{3}{5} >.$$

(b) $\kappa(t) = \dfrac{|\mathbf{r}'(t) \times \mathbf{r}''(t)|}{|\mathbf{r}'(t)|^3} =$

$$\frac{| < -3\sin(t), 3\cos(t), 4 > \times < -3\cos(t), -3\sin(t), 0 > |}{125} =$$

$$\frac{|\det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ -3\sin(t) & 3\cos(t) & 4 \\ -3\cos(t) & -3\sin(t) & 0 \end{bmatrix} |}{125} = \frac{| < 12\sin(t), -12\cos(t), 9 > |}{125} = \frac{15}{125} = \frac{3}{25}.$$

Note that though this curvature is constant, in general, one typically gets different curvatures at different points along a curve.

(c) The point $(-3, 0, 4\pi)$ occurs on the curve $\mathbf{r}(t)$ when $t = \pi$. So, we plug π into $\mathbf{r}'(t)$ to obtain $\mathbf{r}'(\pi) =< 0, -3, 4 >$. Thus, the normal plane is a plane perpendicular to $< 0, -3, 4 >$ and containing point $< -3, 0, 4\pi >$, which has equation $0(x+2) - 3(y-0) + 4(z - 4\pi) = 0$ or $4z - 3y = 16\pi$.

Since the binormal vector $\mathbf{B}(\pi) =< 0, \dfrac{4}{5}, \dfrac{3}{5} >$ is perpendicular to the osculating plane, the osculating plane is $0(x+2) + 4(y-0) + 3(z - 4\pi) = 0$ or $4y + 3z = 12\pi$.

Since the unit principle unit normal vector is perpendicular to the rectifying plane, and $\mathbf{N}(\pi) =< 1, 0, 0 >$ we have that the rectifying plane is $1(x+2) + 0(y-0) + 0(z - 4\pi) = 0$ or $x = -2$.

Theorem 13.7. *Let* $\boldsymbol{r} : I \to \mathbb{R}^3$ *be a regular curve. Then the acceleration* $\boldsymbol{a}(t) = \boldsymbol{r}''(t) = a_T(t)\boldsymbol{T}(t) + a_N(t)\boldsymbol{N}(t)$, *where* $a_T(t) = v' = \dfrac{\boldsymbol{r}'(t) \cdot \boldsymbol{r}''(t)}{|\boldsymbol{r}'(t)|}$ *and* $a_N(t) = \kappa v^2 = \dfrac{|\boldsymbol{r}'(t) \times \boldsymbol{r}''(t)|}{|\boldsymbol{r}'(t)|}$.

Proof. As in the proof of Theorem 13.5, we note $\mathbf{r}''(t) = (\mathbf{T}(t)v(t))' = \mathbf{T}'(t)v(t) + \mathbf{T}(t)v'(t)$. Since $\mathbf{T}'(s) = \kappa\mathbf{N}(s)$, by the Chain Rule we see that $(\mathbf{T}(s(t)))' = \mathbf{T}'(s(t))s'(t) = \mathbf{T}'(t) = \kappa v(t)\mathbf{N}(t)$. Thus, $\mathbf{T}'(t)v(t) = \kappa(v(t))^2\mathbf{N}(t)$, where $\kappa(v(t))^2 = \dfrac{|\mathbf{r}'(t) \times \mathbf{r}''(t)|}{|\mathbf{r}'(t)|^3}|\mathbf{r}'(t)|^2 = \dfrac{|\mathbf{r}'(t) \times \mathbf{r}''(t)|}{|\mathbf{r}'(t)|}$.

Similarly, we see that $\mathbf{r}'(t) \cdot \mathbf{r}''(t) = \mathbf{r}'(t) \cdot (\mathbf{T}'(t)v(t) + \mathbf{T}(t)v'(t)) = v(t)\mathbf{T}(t) \cdot (\mathbf{T}'(t)v(t) + \mathbf{T}(t)v'(t)) = v(t)v'(t)$. Thus, $v'(t) = \dfrac{\mathbf{r}'(t) \cdot \mathbf{r}''(t)}{|\mathbf{r}'(t)|}$.

Hence, $\mathbf{a}(t) = \mathbf{r}''(t) = a_T(t)\mathbf{T}(t) + a_N(t)\mathbf{N}(t)$, where $a_T(t) = v' = \dfrac{\mathbf{r}'(t)\cdot\mathbf{r}''(t)}{|\mathbf{r}'(t)|}$ and $a_N(t) = \kappa v^2 = \dfrac{|\mathbf{r}'(t)\times\mathbf{r}''(t)|}{|\mathbf{r}'(t)|}$. □

Definition 105

Let $\mathbf{r} : I \to \mathbb{R}^3$ be a regular curve. We refer to $a_T(t) = v' = \dfrac{\mathbf{r}'(t)\cdot\mathbf{r}''(t)}{|\mathbf{r}'(t)|}$ as the *tangential component of acceleration* and to $a_N(t) = \kappa v^2 = \dfrac{|\mathbf{r}'(t)\times\mathbf{r}''(t)|}{|\mathbf{r}'(t)|}$ as the *normal component of acceleration.*

Notice that the acceleration of a curve is the sum of vectors in the direction and in the principal unit normal direction to the curve, both of which are in the osculating plane, making the osculating plane, in this sense, the plane that best fits the motion of the curve locally.

Example 13.8. *Find the tangential and normal components of acceleration to the curve* $\boldsymbol{r}(t) =< t^2, 2t, 1 >$ *at* $t = 1$*.*

Solution. $\mathbf{r}'(t) =< 2t, 2, 0 >$ and $\mathbf{r}''(t) =< 2, 0, 0 >$. Plugging into the formulas at $t = 1$ we have $a_T(t) = \dfrac{< 2, 2, 0 > \cdot < 2, 0, 0 >}{| < 2, 2, 0 > |} = \dfrac{4}{2\sqrt{2}} = \sqrt{2}$ and $a_N(t) = \dfrac{| < 2, 2, 0 > \times < 2, 0, 0 > |}{| < 2, 2, 0 > |}$ $= \dfrac{| < 0, 0, -4 > |}{2\sqrt{2}} = \sqrt{2}$.

□

Note that the tangential and normal components of acceleration are usually not equal; these values just happened to be the same in the preceding example.

Finally, we include a theorem to demonstrate how Frenet formulas can be useful.

Theorem 13.8. *Local canonical form. Let* $\boldsymbol{r} : I \to \mathbb{R}^3$ *be a regular parametrized curve with no singular points of order one where* $0 \in I$*. If we use the coordinate system where* $\boldsymbol{r}(0)$ *is the origin and the direction of the* x*-axis is* $\boldsymbol{T}$*, the direction of the* y *axis is* $\boldsymbol{N}$ *and the direction of the* z *axis is* $\boldsymbol{B}$ *where* $\boldsymbol{r}(s) = (x(s), y(s), z(s))$*, then there are functions* $R_x, R_y, R_z : I \to \mathbb{R}$ *so that*

(i) $x(s) = s - \dfrac{\kappa^2}{6}s^3 + R_x$

(ii) $y(s) = \dfrac{\kappa}{2}s^2 + \dfrac{\kappa'}{6}s^3 + R_y$

(iii) $z(s) = -\dfrac{\kappa\tau}{6}s^3 + R_z$

and $\lim\limits_{s\to 0}\dfrac{R_x}{s^3} = \lim\limits_{s\to 0}\dfrac{R_y}{s^3} = \lim\limits_{s\to 0}\dfrac{R_z}{s^3} = 0$

Proof. By Taylor's Theorem we know that $\mathbf{r}(s) = \mathbf{r}(0)+s\mathbf{r}'(0)+\frac{s^2}{2}\mathbf{r}''(0)+\frac{s^3}{6}\mathbf{r}'''(0)+R$ where $\lim\limits_{R\to 0}\frac{R(s)}{s^3} = 0$ and $R = (R_x, R_y, R_z)$. Since we know that $\mathbf{r}'(0) = t(0)$ and $\mathbf{r}''(0) = \kappa(0)\mathbf{N}(0)$ and $\mathbf{r}'''(0) = \kappa'(0)\mathbf{N}(0) + \kappa(0)\mathbf{N}'(0) = \kappa'(0)\mathbf{N}(0) + -\kappa^2(0)\mathbf{T}(0) - \kappa(0)\tau(0)\mathbf{B}(0)$ we can plug these in to the preceding equation to simplify to $\mathbf{r}(s) - \mathbf{r}(0) = s\mathbf{T} + \frac{s^2}{2}\kappa\mathbf{N} + \frac{s^3}{6}(\kappa'\mathbf{N} - \kappa^2 t - \kappa\tau) + < R_x, R_y, R_z >$. Factoring out $\mathbf{T}, \mathbf{N}, \mathbf{B}$ and combining terms, in the given coordinate system this simplifies to the equations given above. □

Note that we can reparametrize any curve so that a given point is the image of 0 and the interval I contains zero, so this technique can be used fairly generally.

This form of a curve lets you readily see some properties relating to the planes we mentioned. For example, when t is small the x component is largest (having the lowest power term), which means that the curve always crosses through the osculating plane, and likewise the y coordinate is positive near a point, making the curve locally in the direction of the unit normal from the rectifying plane near the point $\mathbf{r}(0)$. These sorts of observations help us to understand the shape of the curve.

This development may not obviously explain why we bothered with the torsion as a separate term. To motivate torsion, we mention that if the curvature is positive on an interval and the torsion is known, and both are differentiable, then these two functions uniquely determine the curve on the interval in question up to a rigid motion. This result is called the Fundamental Theorem of the Local Theory of Curves.

Line Integrals

Definition 106

A *vector field* $\mathbf{F}$ on $\mathbb{R}^2$ is a function $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ (or from $\mathbb{R}^3$ to itself), where the images are the associated vectors corresponding to the points in the images rather than the points themselves (using our convention that a point or vector in Euclidean space are interchangeable depending on context). We say that $\mathbf{F}$ is a *gradient vector field* or a *conservative* vector field if $\mathbf{F} = \nabla f$ for some function $z = f(x, y)$. If C is the image of $\mathbf{r} : [a, b] \to \mathbb{R}^n$, which is a C^1 curve, we say that C is *oriented* from $\mathbf{r}(a) = \mathbf{p}$ to $\mathbf{r}(b) = \mathbf{q}$. Let f be a function taking the image of $\mathbf{r}(t)$ into $\mathbb{R}$. Then we define the *line integral of $f(x, y)$ along C with respect to arc length* to be $\int_C f ds = \int_a^b f(\mathbf{r}(t))|r'(t)|dt$. For a vector field $\mathbf{F}$ we say the *line integral of $\boldsymbol{F}$ along C* is $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot \mathbf{T} ds$. Since $\mathbf{T}(t) = \frac{\mathbf{r}'(t)}{|\mathbf{r}'(t)|}$, we can rewrite this expression as $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt$.

We often express path integrals along vector fields in terms of sums of path integrals with respect to individual variables. We define the integral of function $z = f(x, y)$ along C^1 curve

C with respect to a particular variable, say variable x, to be $\int_C f dx = \int_a^b f(\mathbf{r}(t))x'(t)dt$, and define integration with respect to other variables similarly. Thus, if $F =< P, Q >$ we can represent $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)dt = \int_a^b P(\mathbf{r}(t))x'(t) + Q(\mathbf{r}(t))y'(t)dt = \int_C Pdx + Qdy$. Likewise, if $F =< P, Q, R >$ then $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C Pdx + Qdy + Rdz$.

Line integrals were developed in the early eighteen hundreds and they are helpful in analyzing problems involving fluid flow, gas flow, work, force and electrical current. For example, if each vector assigned to a point represents the velocity of a fluid at that point then the vector field could model the flow of a river. If the vectors indicate force due to something like repulsion by a magnetic field or the effects of forces induced by gravity in Newtonian models (in a relativistic model gravity is technically not a force, though it results in forces being applied by one object against another) then the vector field could indicate the directions and magnitudes of the forces. Integrals of functions can indicate accumulation or loss along a path. For instance, if someone is moving along a path on a road with variable amounts of gravel and is collecting gravel along the way then the line integral could indicate the total amount of gravel collected moving along the path. Alternately, it could represent total charge acquired along a plate or total heat energy absorbed moving along a path or total mass of a wire along a path (thinking of the path as accumulating mass in amounts based on the linear density or mass per unit length along the path).

Example 13.9. *Evaluate* $\int_C y^2 ds$, *where* C *is the path along the curve* $y = \sqrt{4 - x^2}$ *from* $(2, 0)$ *to* $(-2, 0)$.

Solution. First, we parametrize the path C, which is $\mathbf{r}(t) =< 2\cos(t), 2\sin(t) >$ over $0 \leq t \leq \pi$. Then $\int_C y^2 ds = \int_0^\pi 4\sin^2(t)\sqrt{4\sin^2(t) + 4\cos^2(t)}dt = (8)(2)(\frac{1}{2})(\frac{\pi}{2}) = 4\pi$.

□

Example 13.10. *Let* $\boldsymbol{F}(x, y) =< 2y, x^2 >$. *Find the line integral of* $\boldsymbol{F}$ *along the line segment* C *from* $(0, 1)$ *to* $(4, 2)$.

Solution. First, we parametrize the path C, which is $\mathbf{r}(t) =< 4t, 1+t >$ over $0 \leq t \leq 1$. Then $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_0^1 \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)dt = \int_0^1 < 2 + 2t, 16t^2 > \cdot < 4, 1 > dt = \int_0^1 8 + 8t + 16t^2 dt = 8t + 4t^2 + \frac{16t^3}{3}\Big|_0^1 = \frac{52}{3}$.

□

Line integrals (or path integrals) have concrete physical representations. We mentioned a few of those above. For instance, if a function f is non-negative and represents the mass per unit length along smooth curve C then the integral of f along C represents the mass of the curve. Another visible way of looking at a line integral of a function is that if the curve is in $\mathbb{R}^2$ then we can think of the function values as indicating the heights of points directly over (or the depths directly below) the curve in a third dimension (like a curtain of variable height). The area of the resulting curtain would be another way to visualize this scalar line integral.

If a vector field $\mathbf{F}$ represents the force acting on a particle as it traverses the curve C then the integral of $\mathbf{F}$ along C represents the work done by the force field $\mathbf{F}$ as the particle moves along the curve C.

Theorem 13.9. *Fundamental Theorem for Line Integrals. Let C be the C^1 path $\boldsymbol{r}(t)$, $a \le t \le b$ in $\mathbb{R}^n$, and let f be a C^1 function on an open set U containing C. Then $\int_C \nabla f \cdot d\boldsymbol{r} = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a))$.*

Furthermore, if Q, parametrized by $\boldsymbol{w} : [a,b] \to \mathbb{R}^n$, is a piecewise smooth path from a point $\boldsymbol{s}$ to a point $\boldsymbol{e}$ that lies within U then $\int_Q \nabla f \cdot d\boldsymbol{r} = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a))$.

Proof. We know $\int_C \nabla f \cdot d\mathbf{r} = \int_a^b \nabla f(\mathbf{r}(t)) \cdot \mathbf{r}'(t)dt = \int_a^b (f(\mathbf{r}(t)))'dt$ using the Chain Rule. But then, by the Fundamental Theorem of Calculus, this is equal to $f(\mathbf{r}(b)) - f(\mathbf{r}(a))$.

Since Q is a piecewise smooth path, we know that Q consists of a finite sequence of paths from $\mathbf{s}$ to some point $\mathbf{x}_1$ followed by a path from $\mathbf{x}_1$ to some point $\mathbf{x}_2$ and so on until we finish with a path from a point $\mathbf{x}_m$ to $\mathbf{e}$, the ending point of the path. Since smooth paths are C^1 we can see from the argument above that $\int_Q \nabla f \cdot d\mathbf{r} = (f(\mathbf{x}_1) - f(\mathbf{s}) + (f(\mathbf{x}_2 - f(\mathbf{x}_1)) + ... + (f(\mathbf{e} - f(\mathbf{x}_m)) = f(\mathbf{e}) - f(\mathbf{s})$ as desired.

□

Example 13.11. *Let C be the path consisting of the line segment from $(0,0,0)$ to $(1,6,2)$ followed by the line segment from $(1,6,2)$ to $(8,0,4)$, followed by the line segment from $(8,0,4)$ to $(1,1,3)$. Let $\boldsymbol{F}(x,y,z) =< 2xe^y, x^2e^y, 2z >$. Find $\int_C \boldsymbol{F} \cdot d\boldsymbol{r}$.*

Solution. First, if $F = \nabla f$ for some potential function f then we know that $f_x = 2xe^y$, $f_y = x^2e^y$, and $f_z = 2z$. Thus, it must follow that $f = x^2e^y + g_1(y,z)$ and $f = x^2e^y + g_2(x,z)$ and $f = z^2 + g_3(x,y)$ for some differentiable g_1, g_2, g_3. We see that setting $g_1(y,z) = g_2(x,z) = z^2$ and $g_3(x,y) = x^2e^y$ we would have the function $f(x,y,z) = x^2e^y + z^2$ in each case, which makes this a potential function for $\mathbf{F}$. From the Fundamental Theorem of Line Integrals it follows that $\int_C \mathbf{F} \cdot \mathbf{dr} = f(1,1,3) - f(0,0,0) = e + 9$.

□

Definition 107

Let $\mathbf{F}$ be a continuous vector field on an open, connected set D containing a piecewise smooth path C which begins at point $\mathbf{a}$ and ends at $\mathbf{b}$. Then we say that $\int_C \mathbf{F} \cdot d\mathbf{r}$ is *path independent* in D if for any other path C_1 which begins and ends at the same points as path C, the line integral $\int_{C_1} \mathbf{F}\cdot\mathbf{dr} = \int_C \mathbf{F}\cdot\mathbf{dr}$, in which case we also write the integral as $\int_{\mathbf{a}}^{\mathbf{b}} \mathbf{F}\cdot\mathbf{dr}$ (since the integral value is only a function of the end points and not the path).

Theorem 13.10. *(a) Let* $\boldsymbol{F}$ *be continuous on an open connected region* D *in* $\mathbb{R}^n$ *and let* $\int_C \boldsymbol{F}\cdot\boldsymbol{dr}$ *be path independent in* D*. Then* $F = \nabla f$ *for some* C^1 *function* f*.*

(b) Let $\boldsymbol{F}$ *be continuous on an open connected region* D*. Then* $\boldsymbol{F}$ *is path independent if and only if* $\boldsymbol{F}$ *is a gradient vector field.*

Proof. (a) We first give the proof assuming $D \subseteq \mathbb{R}^2$ and $\mathbf{F} =< P, Q >$, where P, Q are continuous functions from D into $\mathbb{R}$. Let $(a,b) \in D$ and set $f(x,y) = \int_{(a,b)}^{(x,y)} \mathbf{F}\cdot\mathbf{dr}$ and choose an open ball B so that $(x,y) \in B \subset D$ with $(x_1,y) \in B$ for some $x_1 < x$. Let C_1 be a piecewise smooth path from (a,b) to (x_1,y) and let C_2 be the horizontal line segment from (x_1,y) to (x,y). Then $\int_{(a,b)}^{(x,y)} \mathbf{F}\cdot\mathbf{dr} = \int_{C_1} \mathbf{F}\cdot\mathbf{dr} + \int_{C_2} \mathbf{F}\cdot\mathbf{dr}$. The first of these path integrals is a constant that does not depend on x and along the second path the variable y is constant. Hence, we have $f_x(x,y) = 0 + \frac{\partial}{\partial x}\int_{C_2} P(\mathbf{r}(t))x'(t) + Q(\mathbf{r}(t))y'(t)dt = \frac{\partial}{\partial x}\int_{C_2} P(\mathbf{r}(t))x'(t)dt$. Using the parametrization $\mathbf{r}(t) = (t,y)$, $x_1 \le t \le y$ for C_2 this simplifies to $\frac{\partial}{\partial x}\int_{x_1}^{x} P(t,y)dt = P(x,y)$ by the Fundamental Theorem of Calculus.

To show that $f_y = Q$ you instead choose $(x,y_1) \in B$ with $y_1 < y$, and then the the argument is similar. You just set C_1 to be a path from (a,b) to (x,y_1), and set C_2 to be the path $\mathbf{r}(t) = (x,t)$, $y_1 \le t \le y$ and get that $f_y(x,y) = 0 + \frac{\partial}{\partial y}\int_{C_2} Q(\mathbf{r}(t))y'(t) + 0dt = \frac{\partial}{\partial y}\int_{y_1}^{y} Q(\mathbf{r}(t))dt = Q(x,y)$.

The argument is similar in $\mathbb{R}^n$, letting $\mathbf{a} \in D$ and $f(\mathbf{x}) = \int_{\mathbf{a}}^{\mathbf{x}} \mathbf{F}\cdot\mathbf{dr}$ except that if $D \subseteq \mathbb{R}^n$ we create an additional path for each coordinate derivative, and all the points we have thus far listed have n coordinate entries instead of two. More specifically, we take ball $B_\epsilon(\mathbf{x}) \subset D$ and let C_1 be a path from $\mathbf{a}$ to $(x_1, x_2, ..., x_i - \delta, x_{i+1}, ..., x_n) = \mathbf{x}_0$, where $\delta < \epsilon$. Let C_2 be the line segment parametrized by $\mathbf{r}(t) = (x_1, x_2, ..., t, x_{i+1}, ..., x_n)$, where $x_i - \delta \le t \le x_i$. We can extend this path so that $\mathbf{r}$ is defined on $(x_i - \delta, x_i + \delta)$ so that we can differentiate the path integral at $x_i = t$. When differentiating with respect to x_i we can treat the other variables as fixed and for points $\mathbf{z} = (x_1, x_3, x_3, ..., u, x_{i+1}, ..., x_n)$ where $|u - x_i| < \delta$ we have $f(\mathbf{z}) = \int_{C_1} \mathbf{F}\cdot\mathbf{dr} + \int_{C_2(\mathbf{z})} \mathbf{F}\cdot\mathbf{dr}$, where $C_2(\mathbf{z})$ is the path $(x_1, x_2, ..., t, x_{i+1}, ..., x_n)$ for $x_i - \delta \le t \le u$. Then $\frac{\partial f}{\partial x_i} = (\int_{C_1} \mathbf{F}\cdot\mathbf{dr} + \int_{C_2} \mathbf{F}\cdot\mathbf{dr})_{x_i}$. The first integral is constant and has derivative zero. The second is $\frac{\partial}{\partial x_i}\int_0^u \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t)dt = \mathbf{F}_i(\mathbf{r}(u))$. Thus, at $u = x_i$, the partial derivative is $\mathbf{F}_i(\mathbf{x})$ as desired.

(b) If $\mathbf{F} = \nabla f$, a continuous vector field, then by the Fundamental Theorem of Line Integrals we know that for any piecewise smooth path C from $\mathbf{a}$ to $\mathbf{x}$, the path integral $\int_C \mathbf{F} \cdot \mathbf{dr} = f(\mathbf{x}) - f(\mathbf{a})$, meaning that $\mathbf{F}$ is path independent. Thus, by part (a) it follows that $\mathbf{F}$ is path independent if and only if $\mathbf{F}$ is a gradient vector field. □

From this theorem, we see that path independence and a vector field being conservative or a gradient vector field on a connected open set are all equivalent assuming a vector field's component entries are continuous.

Example 13.12. *Let $\boldsymbol{F}(x, y, z) =< 2xe^{y^2}, 2yx^2e^{y^2}, 2z+x >$ be the force acting on a particle at point (x, y, z) in Newtons. Let C be the path $\boldsymbol{r}(t) =< t, -2t, 2t >$ on $0 \leq t \leq 1$ from $(0, 0, 0)$ to $(1, -2, 2)$. Find the work done by the vector field acting on a particle traversing the path C, assuming the distances indicated by the variables are measured in meters.*

Solution. The work is the integral of the vector fiels along C, which is $\int_C \mathbf{F} \cdot \mathbf{dr}$. Initially, we might hope this vector field is conservative. If $\mathbf{F} = \nabla f$ then antidifferentiating the components of $\mathbf{F}$ should give f. Thus, $f = x^2e^{y^2} + g_1(y, z)$ and $f = x^2e^{y^2} + g_2(x, z)$ and $f = z^2 + xz + g_3(x, y)$ for some differentiable g_1, g_2, g_3. This could almost be done in the sense that if we set $f = x^2e^{y^2} + z^2$ then $\nabla f =< 2xe^{y^2}, 2yx^2e^{y^2}, 2z >= \mathbf{F}_1$. However, we have a leftover zx summand in the third coordinate that cannot be compensated for. There is no way to just add a function of y and z in the $f = x^2e^{y^2} + g_1(y, z)$ and get xz becacuse xz is not a function of only y and z. Thus, this vector field is not conservative. However, we can write $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$, where $\mathbf{F}_2 =< 0, 0, x >$. Since $\int_C \mathbf{F} \cdot \mathbf{dr} = \int_C \mathbf{F}_1 \cdot \mathbf{dr} + \int_C \mathbf{F}_2 \cdot \mathbf{dr}$ we can still use the Fundamental Theorem of Line Integrals to evaluate the first summand, and this will make our work easier. We note that $\int_C \mathbf{F}_1 \cdot \mathbf{dr} = f(1, -2, 3) - f(0, 0, 0) = e^2 + 4$. We use the usual formula to take $\int_C \mathbf{F}_2 \cdot \mathbf{dr} = \int_0^1 < 0, 0, t > \cdot < 1, -2, 2 > dt = t^2\Big|_0^1 = 1$. Hence, $\int_C \mathbf{F} \cdot \mathbf{dr} = e^2 + 4 + 1 = e^2 + 5$ Joules. □

Depending on the vector field, this process may not help. In many cases it is faster to just use the formula for a line integral along a vector field directly rather than find a potential function for only part of the vector field. However, if, when looking for a potential function, you find that there isn't one but that potential function is very close, it may in some cases be worth checking whether writing the original vector field in this way might be beneficial.

It may not be obvious from the definition, but if we have two paths and one ends where the next begins then there is a path consisting of the first path followed by the second. It is also true that the parametrization does not affect the path integral (of either type). And reversing the orientation of a vector field integral negates the integral along the path, whereas reversing the orientation of an integral of a scalar function over a path does not affect the integral.

Independence of parametrization:

Theorem 13.11. *Let C be the smooth curve $\boldsymbol{r}\colon [a,b] \to D \subseteq \mathbb{R}^n$, and let $-C$ be the smooth curve $\boldsymbol{r}^- : [a,b] \to D$ of reversed orientation defined by $\boldsymbol{r}^-(t) = \boldsymbol{r}(a+b-t)$. Then:*

(a) Let $f : D \to \mathbb{R}$ be a function so that $\displaystyle\int_C f ds$ exists. Then $\displaystyle\int_{-C} f ds = \int_C f ds$.

(b) Let $\boldsymbol{F}$ be a continuous vector field on D. Then $\displaystyle\int_{-C} \boldsymbol{F}\cdot \boldsymbol{dr} = -\int_C \boldsymbol{F}\cdot \boldsymbol{dr}$.

Proof. (a) By definition we see that $\displaystyle\int_C f ds = \int_a^b f(\mathbf{r}(t))|\mathbf{r}'(t)|dt$. If we use the substitution $u = a+b-t$ then $t = a+b-u$ and $du = -dt$, which means that $\displaystyle\int_a^b f(\mathbf{r}(t))|\mathbf{r}'(t)|dt = -\int_b^a f(\mathbf{r}(a+b-u)|\mathbf{r}'(a+b-u)|du = \int_a^b f(\mathbf{r}(a+b-u)|\mathbf{r}'(a+b-u)|du = \int_{-C} f ds.$

(b) We have $\displaystyle\int_C \mathbf{F}\cdot\mathbf{dr} = \int_a^b \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{r}'(t)dt$. If we use the substitution $u = a+b-t$ then $t = a+b-u$ and $du = -dt$, and which means that $\displaystyle\int_C \mathbf{F}\cdot\mathbf{dr} = -\int_b^a \mathbf{F}(\mathbf{r}(a+b-u))\cdot\mathbf{r}'(a+b-u)du = \int_a^b \mathbf{F}(\mathbf{r}(a+b-u))\cdot\mathbf{r}'(a+b-u)du$. Now, $(\mathbf{r}(a+b-u))' = -\mathbf{r}'(a+b-u)$ by the chain rule, which means that $\displaystyle\int_a^b \mathbf{F}(\mathbf{r}(a+b-u))\cdot\mathbf{r}'(a+b-u)du = -\int_{-C} \mathbf{F}(\mathbf{r}(a+b-u))\cdot\mathbf{r}'(a+b-u)du$. □

Theorem 13.12. *Let $\boldsymbol{r}_1 : [a_1, b_1] \to D$, and let $\boldsymbol{r}_2 : [a_2, b_2] \to D$ be paths. Then there is a path corresponding to $\boldsymbol{r}_1$ followed by $\boldsymbol{r}_2$, meaning that there is a path $\boldsymbol{r}_3 : [a,b] \to D$ so that for some point $c \in [a,b]$ the restriction $\boldsymbol{r}_3|_{[a,c]}$ of $\boldsymbol{r}_3$ to $[a,c]$ is has an image which is the trace of $\boldsymbol{r}_1$, and the restriction $\boldsymbol{r}_3|_{[a,b]}$ to $[c,b]$ has image which is the trace of $\boldsymbol{r}_2$, and the orientation of $\boldsymbol{r}_3|_{[a,c]}$ is the same as that of $\boldsymbol{r}_1$ and the orientation of $\boldsymbol{r}_3|_{[c,b]}$ is the same as the orientation of $\boldsymbol{r}_2$.*

Proof. Let $a = a_1$ and let $c = b_1$ and let $b = b_1 + b_2 - a_2$. We define $\mathbf{r}_3(t) = \mathbf{r}_1(t)$ if $t \in [a_1, b_1]$ and we define $\mathbf{r}_3(t) = \mathbf{r}_2(t + a_2 - b_1)$ if $t \in [b_1, b_1 + b_2 - a_2]$. Then $\mathbf{r}_3||_{[a_1,b_1]} = \mathbf{r}_1$, and for every $a_2 + t \in [a_2, b_2]$ there is exactly a corresponding value $b_1 + t \in [b_1, b_1 + b_2 - a_2]$ so that $\mathbf{r}_3(b_1 + t) = \mathbf{r}_2(b_1 + t + a_2 - b_1) = \mathbf{r}_2(a_2 + t)$ for all $0 \le t \le b_2 - a_2$ and vice versa. Thus, the trace of $\mathbf{r}_2$ and the trace of $\mathbf{r}_3|_{[b_1,b_1+b_2-a_2]}$ are the same. We note that the starting and ending points of $\mathbf{r}_3|_{[b_1,b_1+b_2-a_2]}$ are the starting and ending point of $\mathbf{r}_2$ as well, so the orientations are preserved. □

Theorem 13.13. *Let C_1 be the smooth curve $\boldsymbol{r}_1 : [a_1, b_1] \to D \subseteq \mathbb{R}^n$, and let C_2 be the smooth curve $\boldsymbol{r}_2 : [a_2, b_2] \to D$, where C_1 and C_2 have the same orientation (starting and ending point). If $f : D \to \mathbb{R}$ is continuous and $\boldsymbol{F}$ is a continuous vector field on D then $\displaystyle\int_{C_1} f ds = \int_{C_2} f ds$ and $\displaystyle\int_{C_1} \boldsymbol{F} \cdot \boldsymbol{dr} = \int_{C_2} \boldsymbol{F} \cdot \boldsymbol{dr}$.*

Proof. First, note that the requirement that f and $\mathbf{F}$ and the derivatives $\mathbf{r}_1$ and $\mathbf{r}_2$ be continuous guarantees that all of these integrals exist since all continuous functions are integrable and we know that $f(\mathbf{r}(t))|\mathbf{r}'(t)|$ and $\mathbf{F} \cdot \mathbf{r}'$ are continuous.

Let C be given by $\mathbf{r} : [a, b] \to \mathbb{R}^n$ is a smooth curve then we have shown we can parametrize the curve with respect to arc length as $S(t) = \displaystyle\int_a^t |\mathbf{r}'(u)| du$ defined on $[a, b]$, where $S(t)$ is the length of the curve $\mathbf{r}$ restricted to $[a, t]$. This is a one to one increasing function whose derivative is $|\mathbf{r}'(t)|$, so S is C^1. Hence, the inverse $t(S)$ of $S(t)$ exists, is increasing and is C^1 on $[0, L]$, where L is the length of C. We could then parametrize $\mathbf{r}$ with respect to arc length as C_S, which is $\mathbf{r}(t(S)) : [0, L] \to \mathbb{R}^n$, and the image would be the trace of C (so the trace of C and C_S are the same).

Thus, $\displaystyle\int_C f ds = \int_a^b f(\mathbf{r}(t))|\mathbf{r}'(t)| dt$ and $\displaystyle\int_{C_S} f ds = \int_0^L f(\mathbf{r}(t(S)))|\mathbf{r}(t(S))'| dS = \int_0^L f(\mathbf{r}(t(S)))|\mathbf{r}'(t(S))||t'(S)| dS$
$\displaystyle\int_0^L f(\mathbf{r}(t(S)))|\mathbf{r}'(t(S))|t'(S) dS$ since $t(S)$ is increasing and has positive derivative. Setting $u = t(S)$, we have $du = t'(S) dS$, so the integral becomes $\displaystyle\int_a^b f(\mathbf{r}(u)|\mathbf{r}'(u)| du = \int_C f ds$.

Thus, replacing $\mathbf{r}$ with $\mathbf{r}_1$ and then with $\mathbf{r}_2$, we see that $\displaystyle\int_{C_1} f ds = \int_{C_2} f ds$.

An integral along a vector field is just an integral of the function $f = \mathbf{F} \cdot \mathbf{T}$, and is therefore a particular case of the preceding result, meaning that $\displaystyle\int_{C_1} \mathbf{F} \cdot \mathbf{dr} = \int_{C_2} \mathbf{F} \cdot \mathbf{dr}$.

□

Combining the results above we can see that if C_1 is any smooth path from $\mathbf{a}$ to $\mathbf{b}$ and C_2 is any smooth path with the same trace from $\mathbf{b}$ to $\mathbf{a}$ then $\displaystyle\int_{C_1} \mathbf{F} \cdot \mathbf{dr} = -\int_{C^2} \mathbf{F} \cdot \mathbf{dr}$ for any continuous vector field $\mathbf{F}$, and for any C^1 function f it follows that $\displaystyle\int_{C_1} f ds = \int_{C^2} f ds$.

Flux and Circulation:

Definition 108

Let C be a smooth closed curve parametrized by $\mathbf{r} : [a, b] \to \mathbb{R}^n$ and $\mathbf{F}$ be a continuous vector field. Then $\displaystyle\int_C \mathbf{F} \cdot \mathbf{dr}$ is the *circulation* of $\mathbf{F}$ along C. If C is in $\mathbb{R}^2$ and $\mathbf{r}'(t) =< x'(t), y'(t) >$ and $\mathbf{F} =< P, Q >$, and $\mathbf{n}(t) = \dfrac{< y'(t), -x'(t) >}{\sqrt{(x'(t))^2 + (y'(t))^2}}$

then the *flux* through **F** in direction $\mathbf{n}(t)$ is $\int_a^b \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{n}(t)ds = \int_C -Qdx + Pdy$.
Alternately, if $\mathbf{n}(t) = \frac{<-y'(t), x'(t)>}{\sqrt{(x'(t))^2+(y'(t))^2}}$ then the flux through **F** in direction $\mathbf{n}(t)$ is $\int_a^b \mathbf{F}(\mathbf{r}(t))\cdot\mathbf{n}(t)ds = \int_C Qdx - Pdy$.

If the vector field F represents velocity of a fluid then the circulation represents the what the amount of fluid flowing around a curve per unit of time approaches as the fluid approaches velocity as it is along the line. The circulation is not actually the quantity of water flowing around the curve per unit time, however, because **F** will vary away from the curve. The units are distance times distance per unit time (velocity) however, so it is measured in square distance units per time unit. If, instead of a curve, we had a tunnel around the curve of a diameter to admit one square unit of the fluid and there was no compression in the fluid due to the curvature (imagine the tunnel is in the fourth dimension) then the circulation could indicate the actual amount of fluid per unit time that is moving along the closed curve if the velocity of all water in a perpendicular cross section to the curve were the listed vector field values. Instead, it represents the rate a volume of fluid passing around the curve for fluid in a tube of small radius divided by the area of a perpendicular cross section of the tube approaches as the radius of the tube approaches zero. It is an idealized notion of rate of fluid flow which is only locally accurate.

Flux through a curve is only defined in $\mathbf{R}^2$ (flux through a surface will be defined in $\mathbb{R}^3$ later). We think of a curve as having two normal directions. If **F** represents the velocity of a fluid then the flux through the curve represents the net amount of fluid passing through the curve per unit time in the direction of the indicated normal vector **n** (the fluid passing from the side of the curve that **n** is pointing away from to the side of the curve that **n** is pointing towards). Unlike circulation, flux measures the area of fluid passing through the curve per unit time (it is not a limit of such rates for fluid very near the curve).

Example 13.13. *Let C be the circle $x^2+y^2=1$ traversed once counterclockwise starting at the point $(1,0)$. Let $\boldsymbol{F} =< 2y, y+1 >$ be the velocity of the fluid. Assuming variable distances are in meters and velocity is in meters per second, find:*

(a) The circulation around C.

(b) The flux of $\boldsymbol{F}$ passing outwards through C.

Solution. Let C be parametrized by $\mathbf{r}(t) =< \cos(t), \sin(t) >$.

(a) We check that **F** is not conservative. Antidifferentiating both coordinates, if $\nabla f = \mathbf{F}$ we would need $f = 2yx + g_1(y)$ and $f = xy + x + g_2(x)$. There is no $g_2(x)$ we could add to add another xy so F is not conservative. We will just go directly to the definition to get $\int_C \mathbf{F}\cdot\mathbf{dr} = \in_0^{2\pi} < 2\sin(t), \sin(t)+1 > \cdot < -\sin(t), \cos(t) > dt = \int_0^{2\pi} -2\sin^2(t) + \sin(t)\cos(t) + \cos(t)dt = -2\pi$ square meters per second is the circulation using Wallis's formula.

(b) Using the parametrization given, $\mathbf{r}'(t) =< -\sin(t), \cos(t) >$ and the unit normal direction that points out through the circle is $< \cos(t), \sin(t) >$. So, the flux is $\int_0^{2\pi} <$

$2\sin(t), \sin(t)+1 > \cdot < \cos(t), \sin(t) > dt = \int_0^{2\pi} 2\sin(t)\cos(t) + \sin^2(t) + \sin(t)dt = \pi$ square meters of fluid per second.

□

The interpretation of a negative value of circulation is that the fluid is moving against the direction of the orientation of the curve. A negative value for flux would indicate that the net fluid flow would be through the curve in the opposite direction to the orientation listed.

Green's Theorem

It should be noted that all simple closed curves separate the plane into two open subsets (the complement of the curve is two disjoint open subsets), one of which is bounded and both of which have boundary equal to the simple closed curve. This result is called the Jordan Curve Theorem, and its proof is messy, so we will only prove Green's Theorem only for some special cases for which the Jordan Curve Theorem is not needed. We only refer to the bounded open set in this separation as the region bounded by the curve. Recall that we have already defined orientation of an arc. We can also assign an orientation to a simple smooth curve (even if it has no end points) which is just a choice of tangent vector (either $\mathbf{T}(t)$ or $-\mathbf{T}(t)$) over the curve (which must be a continuous function of t for a smooth curve). Such choices determine the end and start of a curve if that curve is also a path. We address orientation for smooth closed curves in a similar manner.

Definition 109

If C is a smooth curve than an orientation of C is a C^1 function $\mathbf{O} : C \to \mathbb{R}^n$ assigning a vector to each point $\mathbf{p}$ of C a tangent vector to C at $\mathbf{p}$ of unit length. If $\mathbf{r} : I \to \mathbb{R}^n$ is a smooth curve with trace C then the orientation induced by $\mathbf{r}$ is the function $\mathbf{O}(\mathbf{r}(t)) = \mathbf{T}(t)$ for all $t \in I$. If C is the piecewise-smooth closed curve $\mathbf{r} : [a, b] \to \mathbb{R}^2$ whose trace is the boundary of the bounded open set E, then we say that C is *positively oriented* or oriented *counterclockwise* if for every point $\mathbf{r}(x) \in C$ at which C is differentiable, all points sufficiently close to and to the left of $\mathbf{r}(x)$ are contained in E. More precisely, if $\mathbf{T}(x) =< a, b >$ then there is some $\epsilon > 0$ so that if $0 < t < \epsilon$ then $\mathbf{r}(x) + t < -b, a >\in E$. Likewise, C is *negatively oriented* or oriented *clockwise* if for every point $\mathbf{r}(x) \in C$ at which C is differentiable, all points sufficiently close to and to the right of $\mathbf{r}(x)$ are contained in E. More precisely, if $\mathbf{T}(x) =< a, b >$ then there is some $\epsilon > 0$ so that if $0 < t < \epsilon$ then $\mathbf{r}(x) + t < b, -a >\in E$. If $\mathbf{w} : [a, b] \to \mathbb{R}^n$ is a smooth curve W, and C is a smooth curve which is a subset of W then the orientation of C induced by W is the orientation of C with parametrization $\mathbf{w} : [c, d] \to \mathbb{R}^n$, where $[c, d] = \mathbf{w}^{-1}(C)$.

For a smooth path we can also define an orientation to be a choice of starting and ending point of the path determined by the parametrization. In other words, if $\mathbf{r} : [a, b] \to \mathbb{R}^n$ is a smooth path then its orientation is from $\mathbf{r}(a)$ to $\mathbf{r}(b)$. The corresponding induced orientation is $\mathbf{T}(t)$.

Once a choice of unit tangent vector for a smooth curve is made at one point on the trace of a smooth curve C, that choice determines the choice of unit tangent vector at all other points for a given orientation. The reason is that if we let $\mathbf{O}$ be the orientation of C and $\mathbf{r} : I \to \mathbb{R}^n$ be any parametrization for C, then either $\mathbf{O}(\mathbf{r}(t)) = \mathbf{T}(t)$ or $\mathbf{O}(\mathbf{r}(t)) = -\mathbf{T}(t)$ for every $t \in I$. Hence, if we define $\mathbf{W}(t) = \mathbf{O}(\mathbf{r}(t)) \cdot \mathbf{T}(t)$ then $\mathbf{W}$ is a dot product of two C^1 functions and is C^1 and therefore continuous. Since $\mathbf{W}$ has a range contained in $\{-1, 1\}$, it is only possible for $\mathbf{W}$ to be continuous if $\mathbf{W}$ is constant (because the continuous image of a connected set is connected), so either $\mathbf{W}(t) = 1$ for all $t \in I$, meaning that $\mathbf{O}(\mathbf{r}(t)) = \mathbf{T}(t)$ for all $t \in I$, or $\mathbf{W}(t) = -1$ for all $t \in I$, meaning that $\mathbf{O}(\mathbf{r}(t)) = -\mathbf{T}(t)$ for all $t \in I$. Hence, the choice of induced orientation is determined by the unit tangent vector assigned by that orientation at any given point, there are only two possible orientations for a given smooth curve, and the choice of starting and ending points for a smooth path corresponds to the listed orientation of C.

We first prove Green's Theorem for regions of type I and II. This can be used to prove Green's Theorem for any finite union of regions of type I and II which only intersect along their boundary curves, and this is sufficient to allow us to use Green's Theorem for all of the regions we are interested in addressing in this text.

Theorem 13.14. *Special case of Green's Theorem. Let C be a positively oriented closed path bounding a region E which can be expressed as both the region between $y = g_1(x)$ and $y = g_2(x)$, where $g_1(x) \leq g_2(x)$ over $a \leq x \leq b$, and as the region between $x = h_1(y)$ and $x = h_2(y)$ where $c \leq y \leq d$. Let $\boldsymbol{F} =< P, Q >$ be a C^1 vector field over a connected open set containing $\overline{E}$. Then $\int\int_E \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}$.*

Proof. First, we express C as sequence of four paths $C_1 : \mathbf{r}_1(t) =< t, g_1(t) >$, $a \leq t \leq b$, followed by $C_2 : \mathbf{r}_2(t) =< b, t >$, $g_1(b) \leq t \leq g_2(b)$, followed by $C_3 : \mathbf{r}_3(t) =< -t, g_3(-t) >$, $-b \leq t \leq -a$, followed by $C_4 : \mathbf{r}_4(t) =< a, -t >$, $-g_2(a) \leq t \leq -g_1(a)$.

We will show that $\int\int_E \frac{\partial P}{\partial y} dA = \int_C -P dx$ and $\int\int_E \frac{\partial Q}{\partial x} dA = \int_C Q dy$. First, note that $\int\int_E \frac{\partial P}{\partial y} dA = \int_a^b \int_{g_1(x)}^{g_2(x)} \frac{\partial P}{\partial y} dy dx = \int_a^b P(x, g_2(x)) - P(x, g_1(x)) dx$.

The integral $\int_C P(x,y)dx = \int_{C_1} P(x,y)dx + \int_{C_2} P(x,y)dx + \int_{C_3} P(x,y)dx + \int_{C_4} P(x,y)dx$. We can see that $x'(t) = 0$ along C_2 and C_4 which means that $\int_{C_2} P(x,y)dx = 0 = \int_{C_4} P(x,y)dx$. Also, $\int_{C_1} P(x,y)dx = \int_a^b P(t, g_1(t))x'(t)dt = \int_a^b P(t, g_1(t))dt$ and $\int_{C_3} P(x,y)dx = \int_{-b}^{-a} P(-t, g_2(-t))(-1)dt$. Setting $u = -t$ this integral becomes $\int_b^a P(u, g_2(u))du = -\int_a^b P(t, g_2(t))dt$. Thus, $\int_C -P(x,y)dx = \int_a^b P(t, g_2(t)) - P(t, g_1(t))dt$ as desired.

The process for Q is similar. As before, $\int\int_E \frac{\partial Q}{\partial x} dA = \int_c^d \int_{h_1(x)}^{h_2(x)} \frac{\partial Q}{\partial x} dx dy = \int_c^d Q(h_2(y), y) - Q(h_1(y), y)dy$.

This time we divide C into paths $C_1 : \mathbf{r}_1(t) =< h_1(-t), -t >$, $-d \leq t \leq -c$ followed by $C_2 : \mathbf{r}_2(t) =< t, c >$, $h_1(c) \leq t \leq h_2(c)$ followed by $C_3 : \mathbf{r}_3(t) =< h_2(t), t >$, $c \leq t \leq d$ and finally $C_4 : \mathbf{r}_4(t) =< -t, d >$, $-h_2(d) \leq t \leq -h_1(d)$. As before, $\int_C Q(x,y)dy = \int_{C_1} Q(x,y)dy + \int_{C_2} Q(x,y)dy + \int_{C_3} Q(x,y)dy + \int_{C_4} Q(x,y)dy$. Since $y'(t) = 0$ on paths C_2 and C_4 we know that $\int_{C_2} Q(x,y)dy = 0 = \int_{C_4} Q(x,y)dy$. Next, $\int_{C_3} Q(x,y)dy$ $= \int_c^d Q(h_2(t), t)(1)dt$ and $\int_{C_1} Q(x,y)dy = \int_{-d}^{-c} Q(h_1(-t), -t)(-1)dt = \int_c^d -Q(h_1(t), t)dt$. Thus, $\int\int_E \frac{\partial Q}{\partial x} dA = \int_C Qdy$ and the proof is complete.

□

Example 13.14. *Find the path* $\int_C \boldsymbol{F} \cdot d\mathbf{r}$ *if* C *is the unit circle traversed once clockwise and* $F =< x^3 + 3y + ye^x, e^x - \cos(y) + x^2 >$.

Solution. Since the orientation of the curve is clockwise, we negate the integral we get from Green's Theorem. If D represents the unit disk then this gives us $-\int\int_D \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \mathbf{F} \cdot d\mathbf{r}$ which is $\int\int_D (3 + e^x) - (e^x + 2x)dA$. Using the symmetry of the unit disk, we see that $\int\int_D 3dA = 3\pi$ (since integrating a constant over an area gives the constant times the area) and $\int\int_D -2xdA = 0$ (since both the unit circle and the function are symmetric about the origin). Alternately, converting this last integral to polar becomes $\int_0^{2\pi}\int_0^1 2r\cos(\theta)drd\theta = \int_0^{2\pi}\cos(\theta)d\theta\int_0^1 rdr = 0$ since $\int_0^{2\pi}\cos(\theta)d\theta = 0$. Thus, the desired path integral is 3π.

□

Definition 110

We will refer to a set $E \subseteq \mathbb{R}^2$ as being *piecewise type one and two*, if it is a union of finitely many regions $R_1, ..., R_k$ of both type one and type two so that for each $i > 1$ it is true that $R_i \cap \bigcup_{j=1}^{i-1} R_j \neq \emptyset$, and is a piecewise smooth simple curve $\mathbf{r} : [a,b] \to \mathbb{R}^2$ which is contained in $\partial(R_i) \cap \partial(\bigcup_{j=1}^{i-1} R_j)$, so that $\mathbf{r}((a,b)) \subset (R_i \cup R_j)^\circ$.

Note that these are not standard terms. We use them in this text because they encapsulate all the regions we are planning on proving Green's Theorem for. This is broad enough class of regions that it will include nearly any curve you are likely to wish to use

Green's Theorem for without specifically going out of your way to find a counterexample. Essentially, a piecewise type one and two region is one where you can keep adding extra type one and two regions on a boundary edge until the union is the entire region.

It is also possible to prove that $\mathbf{r}((a,b))$ would be interior to $R_i \cup R_j$ (we didn't have to make it part of the definition). Thus, parts of the definition are redundant, but because we do not wish to prove these things we have stated a definition with more strict conditions.

It may be worth noting that some statements of Green's Theorem found in calculus texts are excessively optimistic. By reason of the Jordan curve theorem, any simple closed curve is the boundary of a bounded open set E which can then be shown to be connected and simply connected, but we do not wish to prove that theorem in this text because it would be a tangent that would take us too far afield. Some statements of the Jordan curve theorem state that every smooth closed curve bounds a region E so that the conclusion of Green's Theorem holds. However, it is known that we can have a smooth closed curve which is the boundary of a bounded region E in $\mathbb{R}^2$ so that E is not even a Jordan region (so we cannot integrate over E). This was determined by W. F. Osgood in 1903 (Proceedings of the American Mathematical Society, volume 4). Our version of Green's Theorem is not as general as it could be, but the form given below is sufficient for our purposes.

Theorem 13.15. *Let E be a piecewise type one and two region in the uv-plane. Then the boundary of E is a piecewise smooth closed curve C. Let $\boldsymbol{F} =< P,Q >$ be a C^1 vector field over a connected open set containing $\overline{E}$. Then $\displaystyle\int\int_E \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}$ if C is positively oriented.*

Proof. We will proceed inductively, appending regions one at a time.

First, assume that R_1 and R_2 are closed regions whose boundaries are the traces of positively oriented piecewise smooth simple closed curves C_1 defined by $\mathbf{r}_1 : [a_1, b_1] \to \mathbb{R}^2$ and C_2 defined by $\mathbf{r}_2 : [a_2, b_2] \to \mathbb{R}^2$ respectively, having the property that for any C^1 vector field $\mathbf{F} < P, Q >$ it is true that $\displaystyle\int_{C_1} \mathbf{F} \cdot d\mathbf{r} = \int\int_{R_1} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA$ and $\displaystyle\int_{C_2} \mathbf{F} \cdot d\mathbf{r} = \int\int_{R_2} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA$. Assume further that $R_1 \cap R_2 = C$, the trace of piecewise simply smooth curve C_3 defined by $\mathbf{r}_3 : [a_3, b_3] \to \mathbb{R}^2$ and that $R_1 \cup R_2$ is connected and $\mathbf{r}_3((a_3, b_3))$ is in the interior of $R_1 \cup R_2$. Let $E = R_1 \cup R_2$. We wish to show that $\displaystyle\int\int_E \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \mathbf{F} \cdot d\mathbf{r}$.

W know $R_1^\circ \cap R_2^\circ = \emptyset$. We can assume that $\mathbf{r}_3$ is oriented so that the left direction is in the direction of the interior of R_1 (re-labeling which curve is C_1 if necessary). Hence, C_3 is a portion of the positively oriented piecewise smooth boundary C_1 and is oriented in the same direction (meaning the unit tangent directions are the same as in $\mathbf{r}_2$ at each point on the trace). We refer to $\mathbf{r}_3^- : [a_3, b_3] \to \mathbb{R}^2$ as being defined by $\mathbf{r}_3^-(t) = \mathbf{r}_3(a + b - t)$, which has the same trace and opposite orientation and is also a simple piecewise smooth curve. Since the interior of R_1 is left of the unit tangent direction along the curve R_1 it follows that the interior of R_2 is right of that direction, or left of the tangent direction of $\mathbf{r}_3^-$. Hence $\mathbf{r}_3^-([a_3, b_3]) \subseteq C_2$ and has orientation which is the same as that of C_2 on C_3 (meaning the unit tangent directions are the same as in $\mathbf{r}_2$ at each point on the trace).

Next, we note that if $\mathbf{x} \in \partial(R_1 \cup R_2)$ then either $\mathbf{x} \in \partial(R_1)$ or $\mathbf{x} \in \partial(R_2)$ which means that $\mathbf{x} \in C_1 \setminus \mathbf{r}((a_3, b_3))$ or $\mathbf{x} \in C_2 \setminus \mathbf{r}((a_3, b_3))$ since no point of $\mathbf{r}((a_3, b_3))$ is in the boundary of $R_1 \cup R_2$ (all such points are in the interior). Likewise, if $\mathbf{x} \in C_1 \cup C_2 \setminus \mathbf{r}((a_3, b_3))$ then for every $\epsilon > 0$ the ball $B_\epsilon(\mathbf{x})$ contains a point of either R_1 or R_2. If $\mathbf{x} \notin C_1$ then we can choose ϵ small enough so that $B_\epsilon(\mathbf{x}) \cap R_2 = \emptyset$ since R_2 is closed, and we know that $B_\epsilon(\mathbf{x})$ contains a point $\mathbf{p}$ which is not in R_1 and since $\mathbf{p} \notin R_2$ either we see that $\mathbf{x} \in \partial(R_1 \cup R_2)$. Likewise, if $\mathbf{x} \in C_2 \setminus C_1$ then $\mathbf{x} \in \partial(R_1 \cup R_2)$.

By re-parametrizing if necessary, we can choose $\mathbf{r}_1$ so that the images of the end points a_1 and b_1 under $\mathbf{r}_1$ are not contained in $\mathbf{r}_3([a_3, b_3])$. Choose q so that $\mathbf{r}_1(q) = \mathbf{r}_3(a)$. Then for any $\epsilon > 0$ we can find a $\delta > 0$ so that if $x \in (q - \delta, q + \delta)$ then $|\mathbf{r}_1(q) - \mathbf{r}_1(x)| < \epsilon$. Note that q is the first point of $[a_1, b_1]$ whose image is contained in the trace of $\mathbf{r}_3$. Hence, there are points in $C_1 \setminus \mathbf{r}_3((a_3, b_3))$ within any positive distance of $\mathbf{r}_3(a)$ since $\mathbf{r}_3((a_3, b_3))$ does not intersect $\mathbf{r}_1(q - \delta, q)$. Likewise, there are points of $C_1 \setminus \mathbf{r}_3((a_3, b_3))$ that can be chosen within any positive distance of $\mathbf{r}_3(b_3)$. This means that $\mathbf{r}_1(a_3)$ and $\mathbf{r}_3(b_3)$ are limit points of $C_1 \setminus \mathbf{r}_3((a_3, b_3))$ and therefore of $C_1 \cup C_2 \setminus \mathbf{r}_3((a_3, b_3))$. The boundary of a set is always closed, so $\mathbf{r}_1(a_3)$ and $\mathbf{r}_3(b_3)$ are elements of $\partial(R_1 \cup R_2)$. Hence, the boundary or $R_1 \cap R_2$ is exactly $C_1 \cup C_2 \setminus \mathbf{r}_3((a, b))$.

Next, we will show that $C_1 \cup C_2 \setminus \mathbf{r}_3((a, b))$ is a piecewise smooth closed curve. There is a piece decomposition of C_1 consisting of component curves $\mathbf{s}_1 : [a_1 = x_0, x_1] \to \mathbb{R}^2, \mathbf{s}_2 : [x_1, x_2] \to \mathbb{R}^2, ..., \mathbf{s}_2 : [x_{m-1}, x_m = b_1] \to \mathbb{R}^2$. There is some j so that $q \in (x_{j-1}, x_j]$. If we choose w so that $\mathbf{r}_1(w) = \mathbf{r}_3(b_3)$ then there is also some k so that $w \in [x_{k-1}, x_k)$. We let P_1 be the piecewise smooth simple curve whose piece decomposition consists of $\mathbf{s}_1 : [x_0, x_1], \mathbf{s}_2 : [x_1, x_2], ..., \mathbf{s}_2 : [x_{j-1}, q]$. We let P_4 be the piecewise smooth curve whose piece decomposition consists of the component curves $\mathbf{s}_k : [w, x_k], \mathbf{s}_{k+1} : [x_k, x_{k+1}], ..., \mathbf{s}_m : [x_{m-1}, x_m]$. We note that the ending point of P_4 is the same as the starting point of P_1, and $P_1 \cup P_4 = C_1 \setminus \mathbf{r}_3((a_3, b_3))$. Similarly, we can find piecewise smooth curves P_2 and P_3 so that the starting point of P_3 is $\mathbf{r}_3(a_3)$ which is the terminal point of P_1, the ending point of P_2 is the starting point of P_3 and the terminal point of P_3 is the starting point of P_4, which is $\mathbf{r}_3(b_3)$. We note that P_1 intersects P_2 at only its end point and P_4 at only its initial point, and likewise P_2 intersects P_1 at only its initial point and P_3 at only its end point, and P_4 intersects P_3 at only its starting point and P_1 at only its ending point. Thus, P_1 followed by P_2 followed by P_3 followed by P_4 is a positively oriented piecewise smooth closed curve, C_4.

Finally, we note that $\int\int_E \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int\int_{R_1} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA + \int\int_{R_2} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_{C_1} \mathbf{F} \cdot d\mathbf{r} + \int_{C_2} \mathbf{F} \cdot d\mathbf{r} = \int_{P_1} \mathbf{F} \cdot d\mathbf{r} + \int_{P_2} \mathbf{F} \cdot d\mathbf{r} + \int_{P_3} \mathbf{F} \cdot d\mathbf{r} + \int_{P_4} \mathbf{F} \cdot d\mathbf{r} + \int_{C_3} \mathbf{F} \cdot d\mathbf{r} - \int_{C_3} \mathbf{F} \cdot d\mathbf{r} = \int_{C_4} \mathbf{F} \cdot d\mathbf{r}$.

The remainder of the theorem follows inductively. If E is a piecewise type one and two region that is a union of finitely many regions $E_1, ..., E_k$ of both type one and type two so that for each $i > 1$ it is true that $R_i \cap \bigcup_{j=1}^{i-1} E_j \neq \emptyset$, and is a piecewise smooth simple curve $\mathbf{r} : [a, b] \to \mathbb{R}^2$ which is contained in $\partial(E_i) \cap \partial(\bigcup_{j=1}^{i-1} E_j)$, so that $\mathbf{r}((a, b)) \subset (E_i \cup E_j)^\circ$, then we just let R_1 be E_1 and R_2 be E_2 to give us that the boundary of $E_1 \cup E_2$ is a positively oriented piecewise smooth closed curve C, and $\int\int_{E_1 \cup E_2} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \mathbf{F} \cdot d\mathbf{r}$. We then let $R_1 = E_1 \cup E_2$ in the argument above and $R_2 = E_3$ to conclude that the boundary of $E_1 \cup E_2 \cup E_3$ is a positively oriented piecewise smooth closed curve C, and

$\int\int_{E_1 \cup E_2 \cup E_3} \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \mathbf{F} \cdot d\mathbf{r}$. We continue appending additional E_i sets until we have added them all, at which point we conclude that the boundary of E is a positively oriented piecewise smooth closed curve C, and $\int\int_E \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dA = \int_C \mathbf{F} \cdot d\mathbf{r}$.

□

Finding the area enclosed by a smooth closed curve:

One of the uses of Green's Theorem is to find an area $A(R)$ of a piecewise type one and two region R whose boundary is a piecewise smooth positively oriented boundary curve (the curves for which we have proven Green's theorem). If $P(x, y) = -y$ and $Q(x, y) = 0$ and $\mathbf{F}(x, y) =< P, Q >$ then $\int\int_R Q_x - P_y dA = \int_C \mathbf{F} \cdot d\mathbf{r} = \int\int_R 1 dA = A(R)$. Likewise, if $P(x, y) = 0$ and $Q(x, y) = x$ or $P(x, y) = -\frac{y}{2}$ and $Q(x, y) = \frac{x}{2}$ then $\int_C \mathbf{F} \cdot d\mathbf{r} = A(R)$.

Example 13.15. *Use Green's Theorem to show that the region R bounded by the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ has area πab.*

Proof. We can parametrize the ellipse as the curve C defined by $\mathbf{r}(t) =< a\cos(t), b\sin(t) >$ over $0 \leq t \leq 2\pi$. We will set $F =< 0, x >$. Then $Q_x - P_y = 1$, so by Green's Theorem we conclude $A(R) = \int\int_R 1 dA = \int_0^{2\pi} < 0, a\cos(t) >< -a\sin(t), b\cos(t) > dt = \int_0^{2\pi} ab\cos^2(t) dt = \pi ab$ by Wallis's formula. □

Flux through a closed curve:

Green's Theorem provides a shortcut for flux integrals if the flux is to be found through a smooth closed curve bounding a convenient region.

Theorem 13.16. *Let C be the positively oriented smooth closed curve $\boldsymbol{r} : [a, b] \to \mathbb{R}^2$ bounding the piecewise type one and two region E, and let $F =< P, Q >$ be a C^1 vector field on $\mathbb{R}^2$. Let $\boldsymbol{n} = \frac{< y'(t), -x'(t) >}{\sqrt{(x'(t))^2 + (y'(t))^2}}$. Then the flux integral of $\boldsymbol{F}$ through C in direction $\boldsymbol{n}$ is $\int\int_E P_x + Q_y dA$.*

Proof. We note that the flux integral $\int_a^b < P, Q > \cdot < y'(t), -x'(t) > dt = \int_a^b Py'(t) - Qx'(t) dt = \int_C P dy - Q dx = \int\int_E P_x + Q_y dA$ by Green's Theorem. □

Example 13.16. *Let C be the square with sides in the lines $x = \pm 2$ and $y = \pm 2$, bounding region R. Let $\boldsymbol{F} =< 2y, x^2 + 3y >$. Find the flux through C outwards through the square (in the normal direction pointing away from the origin).*

Solution. If C is oriented counterclockwise then the normal direction pointing outwards is the one in which we wish to take the flux integral, which means that $\mathbf{n} = \dfrac{< y'(t), -x'(t) >}{\sqrt{(x'(t))^2 + (y'(t))^2}}$ is the corresponding unit normal direction at every point where the path is differentiable. Hence, the flux is $\displaystyle\int\int_R P_x + Q_y dA = \int\int_R 0 + 3dA = 3(12) = 36.$

□

Divergence and Curl

Divergence and curl are useful features of a vector field which have physical applications as well as being useful in determining whether a three dimensional vector field is conservative. Often, we use notation shortcuts to help is remember divergence and curl as mnemonics. A vector field of any dimension has a divergence which is often written as $\nabla \cdot \mathbf{F}$. The idea behind this notation is to treat ∇ as though it were a vector whose entries are differentiation operators with respect to the variables, and $\mathbf{F}$ as a vector whose entries are the coordinates of the vector field. Multiplication is then replaced by performing the differentation operator on the corresponding coordinate of the vector field. Thus, in three dimensions we would have $\nabla =< \dfrac{\partial}{\partial x}, \dfrac{\partial}{\partial y}, \dfrac{\partial}{\partial z} >$ and $\mathbf{F} =< P, Q, R >$ so that the divergence of the vector field is:

Definition 111

Let $\mathbf{F}$ be a differentiable vector field $< P_1, P_2, ..., P_n >$. The divergence div($\mathbf{F}$) is defined by:

$$\text{div}(\mathbf{F}) = \nabla \cdot \mathbf{F} = \sum_{i=1}^{n} \frac{\partial P_i}{\partial x_i}$$

Of course, this is not actually a dot product, but the notation helps you remember the formula. So, for instance, if $\mathbf{F} =< P, Q >$ then $\mathbf{F} =< P, Q >$ is a two dimensional vector field then we have $\text{div}(\mathbf{F}) = \dfrac{\partial P}{\partial x} + \dfrac{\partial Q}{\partial y}$, and if $\mathbf{F} =< P, Q, R >$ is a three dimensional vector field then $\text{div}(\mathbf{F}) = \dfrac{\partial P}{\partial x} + \dfrac{\partial Q}{\partial y} + \dfrac{\partial R}{\partial Z}$.

A similar mnemonic helps us to remember the curl. In this case, we take a cross product:

Definition 112

Let $\mathbf{F}$ be a differentiable vector field. The curl curl($\mathbf{F}$) is defined by:

$$\operatorname{curl}(\mathbf{F}) = \nabla \times \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ P & Q & R \end{vmatrix} =< R_y - Q_z, P_z - R_x, Q_x - P_y >$$

For a two dimensional vector field $\mathbf{F} =< P, Q >$ we define curl($\mathbf{F}$) $= Q_x - P_y$.

Example 13.17. *Let $\boldsymbol{F} =< x^2, 2xy, ye^z >$. Find the curl and divergence of $\boldsymbol{F}$.*

Solution. First, div($\mathbf{F}$) $= \nabla \cdot \mathbf{F} = 2x + 2x + ye^z = 4x + ye^z$. Next, curl($\mathbf{F}$) $= \nabla \times \mathbf{F} =$
$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ x^2 & 2xy & ye^z \end{vmatrix} =< e^z, 0, 2y >$$

□

It can be helpful to think of the absolute value of divergence as representing net expansion of a vector field away from the point at which the divergence is taken (or compression if the divergence is negative), and magnitude of curl as representing the largest circulation per unit area the vector field spins around a given line, where the curl vector itself points in the direction of that line (and the spin is counter clockwise around that vector as viewed from a point in the direction towards which the curl vector points).

As an intuitive informal description, if you think of a given partial like $\dfrac{\partial P}{\partial x}$ as being positive then that means that the flow in the x direction is getting larger, so assuming the value of the vector field in the x direction is positive, the amount leaving is larger than the amount coming in. We we picture this vector field as representing the velocity of a gas then we are saying that speed at which the gas leaves is greater than the speed at which it enters, so the gas is becoming less dense relative to the x direction. By adding the quantities from all three directions we can get an idea of whether the gas coming in exceeds the gas going out. If the divergence is zero then we say that the vector field is *incompressible* which suggests to the mind the idea that the density is neither being increased nor reduced as one moves along the vector field.

The curl is somewhat harder to picture, but it is possible to show that if C_r is the path around a circle of radius r centered at a point $\mathbf{p} \in \mathbb{R}^3$ in the plane perpendicular to curl($\mathbf{F}$)($\mathbf{p}$) which is counterclockwise as viewed from a point along curl($\mathbf{F}$)($\mathbf{p}$) then $|\operatorname{curl}(\mathbf{F})(\mathbf{p})| = \lim_{r \to 0} \dfrac{1}{\pi r^2} \displaystyle\int_{C_r} \mathbf{F} \cdot d\mathbf{r}$, where $\displaystyle\int_C \mathbf{F} \cdot d\mathbf{r}$ is referred to as the circulation along C_r, so that the $|\operatorname{curl}(\mathbf{F})(\mathbf{p})|$ is the largest circulation per unit area at $\mathbf{p}$. If we were to take any other direction from the point $\mathbf{p}$ and calculate $\lim_{r \to 0} \dfrac{1}{\pi r^2} \displaystyle\int_{C_r} \mathbf{F} \cdot d\mathbf{r}$ then we would get a smaller value than the absolute value of the curl. If the curl of a vector field is the zero vector then we say that the vector field is *irrotational*, meaning that the circulation per unit area (also called circulation density or infinitesimal circulation) is zero.

Another way to determine whether a vector field $\mathbf{F}$ is conservative is by determining whether its curl is zero, which we will discuss after Stokes's Theorem.

Surfaces

Definition 113

Let $\mathbf{r} : E \to \mathbb{R}^3$, where $\mathbf{r}(u,v) =< x(u,v), y(u,v), z(u,v) >$ is a C^1 one to one function whose coordinates' partial derivatives are bounded on E, where E is an open Jordan region in $\mathbb{R}^2$. Let $\mathbf{r}_u =< x_u, y_u, z_u >$ and $\mathbf{r}_v =< x_v, y_v, z_v >$. Let $\mathbf{r}$ also satisfy the property that $\mathbf{r}_u \times \mathbf{r}_v \neq \mathbf{0}$ on E. Then we will refer to $\mathbf{r}$ as a *parametrized surface* (or a parametrization for a surface) and $\mathbf{r}(E)$ as a *surface*. We say that the parametrized surface is *regular* at the point (u,v) if there is some open set $V \subseteq \mathbb{R}^3$ containing $\mathbf{r}(u,v)$ so that $V \cap \mathbf{r}(D) = \mathbf{r}(U)$ for some open set $U \subset D$ containing (u,v), where $\mathbf{r}^{-1} : V \cap \mathbf{r}(D) \to U$ is continuous (in other words, $\mathbf{r}$ restricted to U is a homeomorphism onto $V \cap \mathbf{r}(D)$). If $\mathbf{r}$ is regular at every point of D then we say that $\mathbf{r}$ is a regular parametrized surface and that $\mathbf{r}(D)$ is a *regular surface*.

Let $D \subseteq E$, where $U \subseteq D \subseteq \overline{U} \subset E$ and U is a connected open Jordan region. Then we will refer to $\mathbf{r}(D)$ as a *standard surface* and its parametrization as a *parametrized standard surface* or a parametrization for a standard surface. If $D_1, D_2, ..., D_m$ are non-overlapping sets as described, the interior of whose union is a connected open Jordan region U so that $U \subseteq \bigcup_{i=1}^{m} D_i \subseteq \overline{U} \subset E$, then we refer to $\bigcup_{i=1}^{m} \mathbf{r}_i(D_i)$ as a *piecewise smooth* standard surface. We refer to each $\mathbf{r}_i(D_i)$ as a standard surface *component* of the standard piecewise-smooth surface, and each $\mathbf{r}_i$ as a standard piecewise-smooth surface *component parametrization*.
.

When we refer to a standard surface or a just regular surface $\mathbf{r}(D)$ we mean that $\mathbf{r}$ is a standard (or regular) parametrized surface and that D is a domain satisfying the definition above.

The requirement that $\mathbf{r}^{-1} : V \cap \mathbf{r}(D) \to U$ be continuous in the definition of regular at a point is actually redundant (it will follow automatically if the rest of the definition holds). The usual definition of a regular surface allows for multiple parametrizations to be used rather than just one, covering a surface S with a set of coordinate maps (allowing for messier surfaces), but our definition will be sufficient for the things we wish to do with surfaces.

Another way of saying that a parametrized surface is regular at (u,v) is that $\mathbf{r}(D)$ is locally the graph of a differentiable function near the point $\mathbf{r}(u,v)$, though demonstrating this will require a little bit of work. Thus, the points (u, v) at which $\mathbf{r}$ is regular are the points (u, v) so that $\mathbf{r}(D)$ has a tangent plane and normal vector at $\mathbf{r}(u, v)$. The notions of "standard surface" and regular "at a point" are not standard conventions, but these terms will be helpful in the theorems that follow. It is also worth noting that because of the Extreme Value Theorem, the requirement that the partial derivatives be bounded can be omitted in defining a standard parametrization (because it follows automatically on a compact domain).

Note that the requirement that the cross product remain non-zero is the same as the requirement that one of $\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{vmatrix}$ is non-zero.

Example 13.18. *Let $z = f(x, y)$ be a C^1 function defined on a compact set K containing an open Jordan region D. Let $G = \{(x, y, z) \in \mathbb{R}^3 | (x, y) \in D \text{ and } z = f(x, y)\}$ be the graph of f. Show that the function $\mathbf{r}(u, v) = (u, v, f(u, v))$ on D is a parametrized surface so that the surface $\mathbf{r}(D) = G$ is a (standard) regular surface.*

Proof. Since f is C^1 it follows that the partial derivatives of each component of $\mathbf{r}$ are continuous. Since the derivatives of f are continuous on a compact set K containing D it follows that they are bounded on K (and hence on D) by the Extreme Value Theorem. Since whenever $(x_1, y_1, z_1) \neq (x_2, y_2, z_2)$ in G is is true that $(x_1, y_1) \neq (x_2, y_2)$ it follows that $\mathbf{r}(x_1, y_1) \neq \mathbf{r}(x_2, y_2)$. Thus, $\mathbf{r}$ is one to one. Finally, $\mathbf{r}_u =< 1, 0, f_u >$ and $\mathbf{r}_v =< 0, 1, f_v >$ so the cross product is $\mathbf{r}_u \times \mathbf{r}_v =< -f_u, -f_v, 1 >\neq \mathbf{0}$ on D, which means that $\mathbf{r}$ is a parametrized surface. Finally, to see that $\mathbf{r}(D)$ is regular, for any $(u, v, f(u, v)) \in \mathbf{r}(D)$ we find $r > 0$ so that $E = \overline{B_r(u, v)} \subset D$, so that f is bounded on E by the Extreme Value Theorem so we can find an M so that $M > |f(x, y)|$ on E. Then the set $V = \{(x, y, z) \in \mathbb{R}^3 | (x, y) \in B_r(u, v) \text{ and } -M < z < M\}$ is an open subset of $\mathbb{R}^3$ so that $V \cap \mathbf{r}(D) = \mathbf{r}(B_r(u, v))$. Thus, $\mathbf{r}(D)$ is a regular surface. Since $\mathbf{r}$ is C^1 on K it is also true that $\mathbf{r}(D)$ is a standard surface. □

Note that by the same reasoning if $y = g(x, z)$ (or $x = f(y, z)$ respectively) is a C^1 function on a compact set K containing an open Jordan region D then $\mathbf{r}(u, v) = (u, g(u, v), v)$ (or $\mathbf{r}(u, v) = (g(u, v), u, v)$ respectively) is a parametrized surface whose range is a regular surface.

Let $(u_0, v_0) \in D$, a connected open Jordan region, and let $\mathbf{r}(u, v)$ be parametrized surface $\mathbf{r}(u, v)$. We can find $B_r(u_0, v_0) \subset D$ since D is open. Thus, since $\mathbf{r}$ is continuous, we can find $\delta > 0$ so that if $|(u_0, v_0) - (u, v)| < \delta$ then $\mathbf{r}(u, v) \in B_r(u_0, v_0)$. Thus, we can define $\mathbf{C}_1(t) = \mathbf{r}(u_0 + t, v_0)$, $\mathbf{C}_2(t) = \mathbf{r}(u_0, v_0 + t)$ defined on $(-\delta, \delta)$ which are curves contained in the surface $\mathbf{r}(D)$. Note that $C_1'(0) = \lim_{t\to 0} \frac{\mathbf{r}(u_0 + t, v_0) - \mathbf{r}(u_0, v_0)}{t} = \mathbf{r}_u(u_0, v_0)$, and likewise $C_2'(0) = \mathbf{r}_v(u_0, v_0)$. Since these curves are tangent to the surface $\mathbf{r}(D)$, the vector $\mathbf{r}_u \times \mathbf{r}_v(u_0, v_0)$ is perpendicular to the surface. This assumes that there is a vector perpendicular to the surface, of course, which will be true assuming the surface is locally the graph of a differentiable function, which we will see is true if $\mathbf{r}$ is regular at (u_0, v_0).

We see from this and the example above that the unit normal vector to parametrized surface $\mathbf{r}(u,v)$ at $\mathbf{r}(u_0,v_0)$ is $\dfrac{\mathbf{r}_u \times \mathbf{r}_v(u_0,v_0)}{|\mathbf{r}_u \times \mathbf{r}_v(u_0,v_0)|}$ in general, and that in the specific case of the graph of a function $z = f(x,y)$ the unit normal vector is $\dfrac{(-f_u,-f_v,1)}{\sqrt{1+(f_u)^2+(f_v)^2}}(u_0,v_0)$, assuming these parametrized surfaces are regular at (u_0,v_0).

The fact that parametrized surfaces are one to one means that most surfaces one is likely to think of are standard surfaces. However, a surface could come back in on itself like $S = \{(x,y,z) \in \mathbb{R}^3 | -1 < z < 1$ and either $y = \sin(\dfrac{1}{x})$ and $0 < x < 1$ or $x = 0$ and $-1 < y < 1\}$. We can see that near the origin, there is no open set containing the origin which does not contain infinitely many layers of sheets of curve converging towards a section of the surface containing the origin. This means that there isn't a tangent plane in the sense that we have defined the term, or a normal vector at the origin. Even if we extended the definition to something more general we could alter the curve so that it had slopes that did not converge towards any particular value as they approached the origin so we do have a potential problem defining these ideas at points where the surface is not regular. It is thus worth investigating what implies regularity. We have already demonstrated (in the earlier example) that there is a parametrization for a graph of a function which is regular.

We next note the following which lets us determine continuity of an inverse function for one to one continuous functions on a compact domain.

Theorem 13.17. *Let $K \subset \mathbb{R}^n$ be compact and the $f : E \to \mathbb{R}^t$ be one to one and continuous. Then $f^{-1} : f(E) \to E$ is continuous.*

Proof. Let $\{f(\mathbf{p}_m)\} \to f(\mathbf{p})$ in $f(K)$. Then $\{\mathbf{p}_m\}$ is a bounded sequence (since K is bounded) and has a convergent subsequence $\{\mathbf{p}_{m_i}\}$ which converges to a point $\mathbf{q}$ by the Bolzano-Weierstrass Theorem, where $\mathbf{q} \in K$ since K is closed. Thus, $\{f(\mathbf{p}_{m_i})\} \to f(\mathbf{q})$ since f is continuous. Since $\{f(\mathbf{p}_{m_i})\}$ is a subsequence of $\{f(\mathbf{p}_m)\}$ we know that $\{f(\mathbf{p}_{m_i})\} \to f(\mathbf{p})$. Since f is one to one it follows that $\mathbf{p} = \mathbf{q}$. Let $\epsilon > 0$. Suppose there are infinitely many integers m so that $\mathbf{p}_m \notin B_\epsilon(\mathbf{p})$. Then if we order the correspond sequence members into a subsequence $\{\mathbf{p}_{m_j}\}$ (where $\mathbf{m}_j$ is the jth integer so that $\mathbf{p}_{m_j} \notin B_\epsilon(\mathbf{p})$) then this subsequence has a subsequence $\{\mathbf{p}_{s(m_j)}\}$ which converges to a point $\mathbf{w} \in K \setminus B_\epsilon(\mathbf{p})$ since this set is closed. But then $\{f(\mathbf{p}_{s(m_j)})\} \to f(\mathbf{w})$. This is impossible since $|\mathbf{w} - \mathbf{p}| \geq \epsilon$ and we know that $\{f(\mathbf{p}_{s(m_j)})\} \to f(\mathbf{p})$, which implies that $\mathbf{p} = \mathbf{w}$. Thus, since there are only finitely many integers m so that $\mathbf{p}_m \notin B_\epsilon(\mathbf{p})$, we can choose $k \in \mathbb{N}$ so that if $i \geq k$ then $|\mathbf{p}_i - \mathbf{p}| < \epsilon$, so f^{-1} is continuous at $\mathbf{p}$. □

Using this theorem, it is now possible to see why the condition that $\mathbf{r}^{-1}$ is redundant in the definition of regular at a point because if $\mathbf{r}$ is continuous on and open U then we can find $\epsilon > 0$ so that $\overline{B_\epsilon(u,v)} \subset U$, so $\mathbf{r}$ is continuous on $\overline{B_\epsilon(u,v)}$, so $\mathbf{r}^{-1}$ is continuous on $\mathbf{r}(\overline{B_\epsilon(u,v)})$ which means that r^{-1} is also continuous on $\mathbf{r}(\overline{B_\epsilon(u,v)})$. If $V = \mathbf{r}(D) \cap V$ then by picking $\epsilon_{\mathbf{y}} > 0$ so that $B_{\epsilon_{\mathbf{y}}}(\mathbf{y}) \subset V$ for each $\mathbf{y} \in \mathbf{r}(D)$ we see that $V' = \displaystyle\bigcup_{\mathbf{y} \in \mathbf{r}(D)} B_{\epsilon_{\mathbf{y}}}(\mathbf{y})$ is an open set so that $V' \cap \mathbf{r}(D) = \mathbf{r}(B_\epsilon(u,v))$.

Theorem 13.18. *Let $\boldsymbol{r} : D \to \mathbb{R}^3$ be a parametrized surface on the open Jordan region D, where $(u_0, v_0) \in D$. Then $\boldsymbol{r}$ is regular at (u_0, v_0) if and only if $\boldsymbol{r}(D)$ is a surface which is locally the graph of a C^1 function at $\boldsymbol{r}(u_0, v_0)$.*

Proof. First, assume that $\mathbf{r}(D) = S$ is a surface so that $\mathbf{r}$ is locally the graph of a C^1 function at (u_0, v_0). We will assume that the function is of the form $z = f(x, y)$ (the other cases are similar). Then if $\mathbf{r}(u_0, v_0) = (x_0, y_0, z_0)$ there is a C^1 function $f(x, y)$ on some open ball $B_\epsilon(x_0, y_0)$ so that there is some open set V so that $V \cap \mathbf{r}(D) = G = \{(x, y, z) \in \mathbb{R}^3 | (x, y) \in B_\epsilon(x_0, y_0) \text{ and } z = f(x, y)\}$, the graph of f on $B_\epsilon(x_0, y_0)$. Hence, if we use the parametrization $\mathbf{r}_1(u, v) = (u, v, f(u, v))$ then this is a parametrization which is regular at all points $(u, v) \in B_\epsilon(x_0, y_0)$ (as discussed in the example above). If we set $U = \mathbf{r}^{-1}(\mathbf{r}_1(B_\epsilon(x_0, y_0))$ then U is open since for every $\mathbf{z} \in \mathbf{r}^{-1}(\mathbf{r}_1(B_\epsilon(x_0, y_0))$ there is some $\gamma > 0$ so that $B_\gamma(r_1^{-1}(\mathbf{r}(\mathbf{z})) \subset B_\epsilon(x_0, y_0)$ and there is a $\delta > 0$ so that if $|\mathbf{x} - \mathbf{z}| < \delta$ then $|\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{z})| < \gamma$ which means that $|\mathbf{r}_1^{-1}(\mathbf{r}(\mathbf{z})) - \mathbf{r}_1^{-1}(\mathbf{r}(\mathbf{x}))| < \gamma$ and therefore whenever $\mathbf{x} \in B_\delta(\mathbf{z})$ is is true that $\mathbf{x} \in U$, making U open. It follows that $\mathbf{r}$ is regular at (u_0, v_0).

Next, assume that $\mathbf{r}$ is regular at (u_0, v_0), where $\mathbf{r}(u, v) = (x(u, v)y(u, v), z(u, v))$. Then one of $\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial z}{\partial u} & \dfrac{\partial z}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \\ \dfrac{\partial z}{\partial u} & \dfrac{\partial z}{\partial v} \end{vmatrix}$ is non-zero at (u_0, v_0). We will assume that $\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}$ is non-zero at (u_0, v_0) (the other cases are similar). We then find an open set $U \subseteq \mathbb{R}^2$ containing (u_0, v_0) and an open set V containing $\mathbf{r}(u_0, v_0)$ so that $V \cap \mathbf{r}(D) = \mathbf{r}(U)$ and $\mathbf{r} : U \to V \cap \mathbf{r}(D)$ is a homeomorphism.

If we define $\pi(x, y, z) = (x, y)$ to be the projection onto the xy plane, then $\pi \circ \mathbf{r} = (x(u, v), y(u, v)) : U \to \mathbb{R}^2$ and $\det D(\pi \circ \mathbf{r}) \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} (u_0, v_0) \neq 0$. Hence, by the Inverse Function Theorem, it is true that there are open sets $U_1 = B_\epsilon(u_0, v_0)$ and $V_1 = \pi \circ \mathbf{r}(B_\epsilon(u_0, v_0)$ so that $\pi \circ \mathbf{r}$ is one to one, C^1 and has C^1 inverse g on V_1. We then let $f(x, y) = z(g(x, y))$, where $f : V_1 \to \mathbb{R}$. Since g and z are C^1 we know that f is C^1 on V_1. The graph of f is $G = \{(x, y, z) \in \mathbb{R}^3 | (x, y) \in V_1 \text{ and } z = z(g(x, y))\}$. However, $z(g(x, y))$ is the unique z value corresponding to the point $(u, v) = g^{-1}(x, y) \in U_1$ so that $((x(u, v), y(u, v), z(u, v)) \in \mathbf{r}(U_1)$, which means that $(x, y, z) \in G$ if and only if $(x(u, v), y(u, v), z(u, v)) \in \mathbf{r}(U_1)$, so the image of U_1 under $\mathbf{r}$ is the graph of a differentiable function f. If we define $W = V \cap \pi^{-1}(V_1)$ then $\mathbf{r}(D) \cap W = G$, so $\mathbf{r}(D)$ is locally the graph of a C^1 function near $\mathbf{r}(u_0, v_0)$ as desired.

□

An almost immediate consequence of the preceding theorem is the following:

Theorem 13.19. *Let $F : U \to \mathbb{R}$, where U is open in $\mathbb{R}^3$ containing $\boldsymbol{p} = (x_0, y_0, z_0)$, F is C^1 and $F(x_0, y_0, z_0) = k$. Let $\nabla F(\boldsymbol{x}) \neq 0$ for every $\boldsymbol{x} \in U$ so that $F(\boldsymbol{x}) = k$. Let $S = \{(x, y, z) \in U | F(x, y, z) = k\} = F^{-1}(k)$ (the graph of the relationship $F(x, y, z) = k$) and let $\boldsymbol{r} : D \to \mathbb{R}^3$ be a parametrized surface where $\boldsymbol{r}(D) = S$. Then S is a regular surface.*

Proof. By Theorem 11.28, S is locally the graph of a C^1 function, which means that S is a regular surface by Theorem 13.18. □

Let $\mathbf{r} : D \to \mathbb{R}^3$ be a regular parametrized surface $\mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v))$ in $\mathbb{R}^3$. Recall from Theorem 10.4 that the area of the parallelogram with vector sides $\mathbf{r}_u, \mathbf{r}_v$ is $|\mathbf{r}_u \times \mathbf{r}_v|$. If we were to take the edges of a rectangle $R_{=}[u_0, u_0 + \Delta u] \times [v_0, v_0 + \Delta v] \subset D$ with vector sides $< \Delta u, 0 >$ and $< 0, \Delta v >$ then if the derivative of the transformation ϕ were a constant matrix D whose columns are $\mathbf{r}_u, \mathbf{r}_v$ on R, we would have that the function $\phi(u_0 + h_1, v_0 + h_2) - \phi(u_0, v_0) = (\frac{\partial x}{\partial u}h_1 + \frac{\partial x}{\partial v}h_2, \frac{\partial y}{\partial u}h_1 + \frac{\partial y}{\partial v}h_2, \frac{\partial z}{\partial u}h_1 + \frac{\partial z}{\partial v}h_2)$ (by the Mean Value Theorem for Real Valued Functions on each component function). If we only look at points on the rectangle then $0 \leq h_1 \leq \Delta u$ and $0 \leq h_2 \leq \Delta v$ then the image of the rectangle R (which is, by definition, $\{(u_0 + t_1\Delta u, v_0 + t_2\Delta v)|0 \leq t_1 \leq 1$ and $0 \leq t_2 \leq 1\}$) under $\mathbf{r}$ would be the parallelogram $P = \{(x(u_0, v_0) + \frac{\partial x}{\partial u}\Delta u t_1 + \frac{\partial x}{\partial v}\Delta v t_2, y(u_0, v_0) + \frac{\partial y}{\partial u}\Delta u t_1 + \frac{\partial y}{\partial v}\Delta v t_2, z(u_0, v_0) + \frac{\partial z}{\partial u}\Delta u t_1 + \frac{\partial z}{\partial v}\Delta v t_2)|0 \leq t_1 \leq 1$ and $0 \leq t_2 \leq 1\}$. In other words, the image of the rectangle would be a parallelogram with sides $\mathbf{r}_u\Delta u$ and $\mathbf{r}_v\Delta v$ with area $|\mathbf{r}_u \times \mathbf{r}_v|\Delta u\Delta v$. This means that it is reasonable to think of $\sum_{i=1}^{n}\sum_{j=1}^{m}|\mathbf{r}_u \times \mathbf{r}_v|\Delta u\Delta v$ as approximating the surface area over a rectangle if the function $\mathbf{r}(u, v)$ is continuously differentiable because then for small rectangles the derivative is approximately constant over the rectangle. Much as we defined a line integral with respect to arc length as the limit of a sum of function values times lengths on segments of a curve, we define a scalar surface integral of a function f on a parametrized (or compact paramatrized) surface $S_1 = \mathbf{r}(D)$ as $\lim_{n\to\infty}\lim_{m\to\infty}\sum_{i=1}^{n}\sum_{j=1}^{m} f(\mathbf{r}(u_i^*, v_j^*))|\mathbf{r}_u \times \mathbf{r}_v|\Delta u\Delta v$. This motivates us to define the surface area and surface integral of a standard piecewise-smooth surface $\mathbf{r}(D)$ as follows.

Definition 114

Let $\mathbf{r}(D) = S_1$ be a standard surface. The *surface integral* (or scalar surface integral) of f over S_1 is $\int\int_{S_1} f dS = \int\int_D f(\mathbf{r}(u, v))|\mathbf{r}_u \times \mathbf{r}_v|dA$. The *surface area* of D is $\int\int_{S_1} 1dS = \int\int_D |\mathbf{r}_u \times \mathbf{r}_v|dA$. The surface area or surface integral of a standard piecewise-smooth surface is the sum of the surface areas or integrals of the standard surface components of the surface.

In the case where S_1 is the graph of a C^1 function $z = g(x, y)$ over a domain D we can use the parametrization $\mathbf{r}(u, v) = (u, v, g(u, v))$ on D, on which we have (using the formula listed above) a simplification of the surface area formula:

$$A = \int\int_D \sqrt{1 + g_x^2 + g_y^2}dA$$

The corresponding surface integral formula is:

$$\int\int_{S_1} f dS = \int\int_D f(x, y, g(x, y))\sqrt{1 + g_x^2 + g_y^2} dA$$

Scalar surface integrals give a total quantity over a surface for which an amount per unit area is known along the surface. For instance, integrating mass per unit area over a surface would give the total mass of the surface as the surface integral.

Example 13.19. *Find the surface area of the surface paramatrized by* $\mathbf{r}(u, v) =< u\cos v, u\sin v, v >$, $0 \le u \le 1$, $0 \le v \le \pi$.

Solution. Recall surface area of S_1 is $\iint_D |\mathbf{r}_u \times \mathbf{r}_v| dA$. In this case, $\mathbf{r}_u =< \cos v, \sin(v), 0 >$ and $\mathbf{r}_v =< -u\sin v, u\cos v, 1 >$, so $\mathbf{r}_u \times \mathbf{r}_v = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \cos v & \sin(v) & 0 \\ -u\sin v & u\cos v & 1 \end{bmatrix} =< \sin(v), -\cos(v), u >$ and $|\mathbf{r}_u \times \mathbf{r}_v| = \sqrt{u^2 + 1}$. Thus, $\text{Area}(S_1) = \iint_D |\mathbf{r}_u \times \mathbf{r}_v| dA = \int_0^\pi \int_0^1 \sqrt{u^2 + 1} du dv$. Setting $u = \tan\theta$ we have $du = \sec^2\theta d\theta$, and thus the integral simplifies to $\pi \int_0^{\frac{\pi}{4}} \sec^3\theta d\theta = \frac{\pi}{2}(\sec(\theta)\tan(\theta) + \ln|\sec(\theta) + \tan(\theta)|)\Big|_0^{\frac{\pi}{4}} = \frac{\pi}{2}(\sqrt{2} + \ln(\sqrt{2} + 1))$.

□

Example 13.20. *Find the surface area of the portion of the plane* $z + 2x + 3y = 6$ *in the first octant (where* $x \ge 0$, $y \ge 0$ *and* $z \ge 0$*).*

Solution. First, this is a triangular disk so we could just find the length of its base and its height, but we will use the methods described above instead. We solve for $z = g(x, y) = 6 - 2x - 3y$ to get the graph which gives the desired surface, and use the formula $A = \int\int_D \sqrt{1 + g_x^2 + g_y^2} dA$. When $x = z = 0$ we have $y = 2$ and when $y = z = 0$ we have $x = 3$, so the surface described is the graph of $z = g(x, y)$ over the trianglular disk T given by $0 \le x \le 3$ and $0 \le y \le 2 - \frac{2}{3}x$. This means that the surface area is $\int\int_T \sqrt{1 + 4 + 9} dA = 3\sqrt{14}$. We did not need to use the bounds because T is a three by two triangle, so the base has area three, and we are integrating a constant function in this case so we just multiply the area of T by the constant to get the integral.

□

Just as we did with line integrals along vector fields we also define integral of vector fields along (or perhaps "through" would be a better word than "along") surfaces which we refer to as flux integrals. To evaluate such an integral we need to establish an orientation for the surface just as we would establish an orientation for a path.

Definition 115

Let S_1 be a standard piecewise-smooth surface. If there is a continuous function $N : S_1 \to S^2$, the sphere $S^2 = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 = 1\}$, then we refer to $N(x, y, z)$ as an *orientation* of the surface S_1.

Some surfaces are not orientable (meaning that there is no orientation function as described). The classic example of such a surface is usually the Mobius band, but we are not going to prove that this surface is not orientable. Essentially, we assign a unit normal vector to every point of the surface. If we can do so in such a way that the assigned unit normal vectors vary continuously over the surface S^1 then that is an orientation. There are only two possible orientations for a connected surface. The idea intuitively with a flux integral is that if $\mathbf{F}(x, y, z) = (P, Q, R)$ represents rate of flow (such as fluid flow) at each point then we would like the flux integral to represent the rate at which the fluid passes through the surface S_1 in the direction in which S^1 is oriented. Thus, we define the flux integral to be the surface integral of the component of F in the unit normal direction $\mathbf{n}$ pointing in the direction of the orientation of the surface. For a given standard parametrized surface $\mathbf{r}(D)$, the only possible orientations are $\dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ and $-\dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ throughout the surface. This is because there are exactly two orthogonal vectors to a surface of length one at each point of a surface which is locally the graph of a differentiable function at a given point, namely those stated, and if $\mathbf{N}$ is an orientation for the surface $\mathbf{r}(D)$ then the function $\mathbf{N}(\mathbf{r}(u, v)) \cdot \dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ defined on D can only take two values (1 or -1). The image of $\mathbf{N}(\mathbf{r}(u, v)) \cdot \dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ is connected since D is connected and the function is continuous, meaning that the image is a single point (either 1 or -1), so either $\mathbf{N}(\mathbf{r}(u, v)) = \dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ for all $(u, v) \in D$ or $\mathbf{N}(\mathbf{r}(u, v)) = -\dfrac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}$ for all $(u, v) \in D$.

Definition 116

The *flux* integral (or *vector surface* integral) of a vector field $\mathbf{F}$ over a parametrized surface $\mathbf{r}(D) = S_1$ to be: $\displaystyle\int\int_{S_1} \mathbf{F} \cdot d\mathbf{S} = \int\int_{S_1} \mathbf{F} \cdot \mathbf{n} dS = \int\int_D \mathbf{F}(\mathbf{r}(u, v)) \cdot \frac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|} |\mathbf{r}_u \times \mathbf{r}_v| dA$ which simplifies to

$$\int\int_{S_1} \mathbf{F} \cdot d\mathbf{S} = \int\int_D \mathbf{F}(\mathbf{r}(u, v)) \cdot \mathbf{r}_u \times \mathbf{r}_v dA$$

where $\mathbf{r}_u \times \mathbf{r}_v$ is the orientation of S_1 (or we negate this integral in the case where $\mathbf{r}_u \times \mathbf{r}_v$ is opposite the direction $\mathbf{n}$ in which S_1 is oriented).

In the case where S_1 is the graph of a function $z = g(x, y)$ oriented upwards over a domain D this formula becomes: $\displaystyle\int\int_{S_1} \mathbf{F} \cdot d\mathbf{S} = \int\int_D (P, Q, R) \cdot (-g_x, -g_y, 1) dA$ since $\mathbf{r}_u \times \mathbf{r}_v = (-g_u, -g_v, 1)$ and $\mathbf{r}(u, v) = (u, v, g(u, v))$ on D, and we can exchange the names of the variables u for x and v for y without altering the value of the integral. Hence, this simplifies to:

$$\int\int_{S_1} \mathbf{F}\cdot d\mathbf{S} = \int\int_D -Pg_x - Qg_y + RdA$$

assuming that the orientation of S_1 is upwards (has a positive z component) on S_1. As usual, we negate the integral if the orientation is the opposite (has a negative z component on S_1).

We have not yet proven that surface integrals are independent of the parametrization chosen to generate the surface, which is important for understanding surface integrals.

Theorem 13.20. *Let $\boldsymbol{r}: E \to \mathbb{R}^3$ and $\boldsymbol{r}_1 : E_1 \to \mathbb{R}^3$ be regular paramatrized surfaces, where E and E_1 are open in $\mathbb{R}^2$.*

(a) Then for each $\boldsymbol{x} \in E$ we can find some $U_{\boldsymbol{x}} \subset E$ so that $\phi = \boldsymbol{r}_1^{-1} \circ \boldsymbol{r} : U_{\boldsymbol{x}} \to \phi(U_{\boldsymbol{x}}) \subset E_1$ is a C^1 homeomorphism with $\triangle_\phi \neq 0$ on $U_{\boldsymbol{x}}$.

(b) Let $\boldsymbol{r} : D \to \mathbb{R}^3$ and $\boldsymbol{r}_1 : D_1 \to \mathbb{R}^3$ be standard parametrized surfaces with $\boldsymbol{r}(D) = \boldsymbol{r}_1(D_1)$, where $U \subseteq D \subseteq \overline{U} \subset E \subset \mathbb{R}^2$ and $U_1 \subseteq D_1 \subseteq \overline{U_1} \subset E_1 \subset \mathbb{R}^2$ for Jordan regions U, D, E, where U, U_1, E and E_1 are open. Let f be a continuous function on $\boldsymbol{r}(D)$.

Then $\displaystyle\int_{\boldsymbol{r}(D)} fdS = \int_{\boldsymbol{r}_1(D_1)} fdS$.

Proof. (a) Let $\mathbf{x} \in E$. We can find a unique $\mathbf{z_x} \in D_1$ so that $\mathbf{r}_1(\mathbf{z_x}) = \mathbf{r}(\mathbf{x})$. We choose positive numbers $\epsilon_\mathbf{x}$ and $\delta_\mathbf{x}$ so that $\mathbf{r}^{-1}$ and $\mathbf{r}_1^{-1}$ restricted to $B_{\epsilon_\mathbf{x}}(\mathbf{r}(\mathbf{x})) \cap \mathbf{r}(E)$ are homeomorphisms, and that $B_{\delta_\mathbf{x}}(\mathbf{x}) \subset \mathbf{r}^{-1}(B_{\epsilon_\mathbf{x}}(\mathbf{r}(\mathbf{x}))$ and $B_{\delta_\mathbf{x}}(\mathbf{z_x}) \subset \mathbf{r}_1^{-1}(B_{\epsilon_\mathbf{x}}(\mathbf{r}(\mathbf{x}))$.

We observe that $\mathbf{r}_1^{-1} \circ \mathbf{r} : B_{\delta_\mathbf{x}}(\mathbf{x}) \to V$ is a homeomorphism, where V is the open set $\mathbf{r}_1^{-1} \circ \mathbf{r}(B_{\delta_\mathbf{x}}(\mathbf{x}))$ in E_1. We note at this point that restricting $\mathbf{r}_1^{-1} \circ \mathbf{r}$ to any open subset of $B_{\delta_\mathbf{x}}(\mathbf{x}))$ would similarly result in a homeomorphism onto its image.

Let $\mathbf{r}_1(u,v) =< x(u,v), y(u,v), z(u,v) >$. We know that one of $\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{vmatrix}$, $\begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} \end{vmatrix}$ is non-zero at each point of E. Without loss of generality, we can assume that $\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \neq 0$ at $\mathbf{r}_1(\mathbf{z_x})$. We let $F : V \times \mathbb{R} \to \mathbb{R}^3$ be defined by $F(u,v,w) =< x(u,v), y(u,v), z(u,v)+w >$. Then $\triangle_F((\mathbf{z_x}, 0)) = \begin{vmatrix} \frac{\partial x}{\partial u}(\mathbf{z_x}) & \frac{\partial x}{\partial v}(\mathbf{z_x}) & 0 \\ \frac{\partial y}{\partial u}(\mathbf{z_x}) & \frac{\partial y}{\partial v}(\mathbf{z_x}) & 0 \\ 0 & 0 & 1 \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial u}(\mathbf{z_x}) & \frac{\partial x}{\partial v}(\mathbf{z_x}) \\ \frac{\partial y}{\partial u}(\mathbf{z_x}) & \frac{\partial y}{\partial v}(\mathbf{z_x}) \end{vmatrix} \neq 0$. Hence, by the Inverse Function Theorem we can find a $0 < \gamma_\mathbf{x} < \delta_\mathbf{x}$ so that $B_{\gamma_\mathbf{x}}(\mathbf{z_x}) \subset V$ and $\triangle_F$ is non-zero on $B_{\gamma_\mathbf{x}}((\mathbf{z_x}, 0))$, $F(B_{\gamma_\mathbf{x}}((\mathbf{z_x}, 0)) = W_\mathbf{x}$ is open, and $F^{-1}(x,y,z) =< \alpha(x,y,z), \beta(x,y,z), z - z(\alpha(x,y,z), \beta(x,y,z)) >$ is C^1 on $W_\mathbf{x}$. This means that α and β have continuous first partial derivatives.

Next, note that $F(\mathbf{p},0)=\mathbf{r}_1(\mathbf{p})$ for all $\mathbf{p}\in B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}})$, so $\mathbf{r}_1^{-1}(x,y,z) =< \alpha(x,y,z),\beta(x,y,z) >$ on $\mathbf{r}_1(B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}))\subset W_{\mathbf{x}}$, which means that $\mathbf{r}_1^{-1}$ is C^1. Hence, $\phi=\mathbf{r}_1^{-1}\circ\mathbf{r}$ is a C^1 homeomorphism on $U_{\mathbf{x}}=\mathbf{r}^{-1}\circ\mathbf{r}_1(B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}))$.

Next, we note that ϕ and ϕ^{-1} are C^1 and $\phi\circ\phi^{-1}$ is the identity, so $D\phi\circ\phi^{-1}(\mathbf{z})=1$ for all $\mathbf{z}\in B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}})$, which means that $\det(D\phi(\phi^{-1}(\mathbf{z}))D\phi^{-1}(\mathbf{z}))=1$, which means that both $\det D\phi(\phi^{-1}(\mathbf{z}))$ and $\det D\phi^{-1}(\mathbf{z})$ are non-zero. Since ϕ is a homeomorphism, each point on $U_{\mathbf{x}}$ is $\phi^{-1}(\mathbf{z})$ for some $\mathbf{z}\in B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}}))$. From this we conclude that $\triangle_{\phi}$ and $\triangle_{\phi^{-1}}$ are non-zero on $U_{\mathbf{x}}$ and $B_{\gamma_{\mathbf{x}}}(\mathbf{z}_{\mathbf{x}})$ respectively.

(b) By (a), for each $\mathbf{x}\in\overline{D}$ we can pick a $\epsilon_{\mathbf{x}}>0$ so that $\phi=\mathbf{r}^{-1}\circ\mathbf{r}_1$ is a C^1 homeomorphism with $\triangle_{\phi}\neq 0$ on $B_{\epsilon_{\mathbf{x}}}(\mathbf{x})$. We know that $\overline{D}$ is compact and that $\mathcal{C}=\{B_{\epsilon_{\mathbf{x}}}(\mathbf{x})|\mathbf{x}\in\overline{D}\}$ is an open cover of $\overline{D}$, so by the Lebesgue number theorem we can find $\delta>0$ so that if Q is a set of diameter less than δ which intersects $\overline{D}$ then Q is a subset of an element of $\mathcal{C}$. Let R be a rectangle containing $\overline{D}$ and let G be a grid on R with $|G|<\dfrac{\delta}{2}$. Then by the Change of Variables Theorem, $\displaystyle\int_{r(D)} f dS=\sum_{R_i\in G}\int_{R_i} f\circ\mathbf{r}|\mathbf{r}_u\times\mathbf{r}_v| =$

$\displaystyle\sum_{R_i\in G}\int_{\phi^{-1}(R_i)} f\circ(\mathbf{r}\circ\phi)|\mathbf{r}_u\times\mathbf{r}_v\circ\phi||\triangle_{\phi}|$. We know that $\mathbf{r}\circ\phi=\mathbf{r}\circ(\mathbf{r}^{-1}\circ\mathbf{r}_1)=\mathbf{r}_1$.

Let $\phi(s,t)=(\alpha(s,t),\beta(s,t))$ for $(s,t)\in\phi^{-1}(R_i)$ and let $\mathbf{r}(u,v) =< x(u,v),y(u,v),z(u,v) >$ and $\mathbf{r}_1(s,t) =< x_1(s,t),y_1(s,t),z_1(s,t) >$. Since $\mathbf{r}\circ\phi=\mathbf{r}_1$, it follows that $x((\alpha(s,t),\beta(s,t)))=x_1(s,t)$, $y((\alpha(s,t),\beta(s,t)))=y_1(s,t)$ and $z((\alpha(s,t),\beta(s,t)))=z_1(s,t)$. From the chain rule we have that $D\mathbf{r}_1(s,t)=D\mathbf{r}\circ\phi(s,t)D\phi(s,t)=\begin{bmatrix}\dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v}\\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v}\\ \dfrac{\partial z}{\partial u} & \dfrac{\partial z}{\partial v}\end{bmatrix}(\phi(s,t))\begin{bmatrix}\dfrac{\partial \alpha}{\partial s} & \dfrac{\partial \alpha}{\partial t}\\ \dfrac{\partial \beta}{\partial s} & \dfrac{\partial \beta}{\partial t}\end{bmatrix}(s,t)$. This gives us:

$$\frac{\partial x_1}{\partial s}(s,t)=\frac{\partial x}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial s}(s,t)+\frac{\partial x}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial s}(s,t)$$
$$\frac{\partial y_1}{\partial s}(s,t)=\frac{\partial y}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial s}(s,t)+\frac{\partial y}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial s}(s,t)$$
$$\frac{\partial z_1}{\partial s}(s,t)=\frac{\partial z}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial s}(s,t)+\frac{\partial z}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial s}(s,t)$$
$$\frac{\partial x_1}{\partial t}(s,t)=\frac{\partial x}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial t}(s,t)+\frac{\partial x}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial s}(s,t)$$
$$\frac{\partial y_1}{\partial t}(s,t)=\frac{\partial y}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial t}(s,t)+\frac{\partial y}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial t}(s,t)$$
$$\frac{\partial z_1}{\partial t}(s,t)=\frac{\partial z}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial t}(s,t)+\frac{\partial z}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial t}(s,t)$$

These six equations can be written as:

$$\frac{\partial\mathbf{r}_1}{\partial s}(s,t)=\frac{\partial\mathbf{r}}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial s}(s,t)+\frac{\partial\mathbf{r}}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial s}(s,t)$$
$$\frac{\partial\mathbf{r}_1}{\partial t}(s,t)=\frac{\partial\mathbf{r}}{\partial u}(\phi(s,t))\frac{\partial\alpha}{\partial t}(s,t)+\frac{\partial\mathbf{r}}{\partial v}(\phi(s,t))\frac{\partial\beta}{\partial t}(s,t).\text{ T}$$

From this, we see that $\mathbf{r}_{1_s}\times\mathbf{r}_{1_t}=(\mathbf{r}_u(\phi(s,t))\dfrac{\partial\alpha}{\partial s}(s,t)+\mathbf{r}_v(\phi(s,t))\dfrac{\partial\beta}{\partial s}(s,t))\times(\mathbf{r}_u(\phi(s,t))\dfrac{\partial\alpha}{\partial t}(s,t)+\mathbf{r}_v(\phi(s,t))\dfrac{\partial\beta}{\partial t}(s,t))=\mathbf{r}_u(\phi(s,t))\times\mathbf{r}_v(\phi(s,t))\dfrac{\partial\alpha}{\partial s}(s,t)\dfrac{\partial\beta}{\partial t}(s,t)-\mathbf{r}_u(\phi(s,t))\times\mathbf{r}_v(\phi(s,t))\dfrac{\partial\alpha}{\partial t}(s,t)\dfrac{\partial\beta}{\partial s}(s,t)=$

$\mathbf{r}_u(\phi(s,t))\times\mathbf{r}_v(\phi(s,t))\det\begin{bmatrix}\dfrac{\partial \alpha}{\partial s} & \dfrac{\partial \alpha}{\partial t}\\ \dfrac{\partial \beta}{\partial s} & \dfrac{\partial \beta}{\partial t}\end{bmatrix}(s,t)$.

Thus, $\sum_{R_i\in G}\int_{\phi^{-1}(R_i)} f\circ(\mathbf{r}\circ\phi)|\mathbf{r}_u\times\mathbf{r}_v\circ\phi||\triangle_\phi| = \sum_{R_i\in G}\int_{\phi^{-1}(R_i)} f\circ\mathbf{r}_1|\mathbf{r}_{1_s}\times\mathbf{r}_{1_t}|$, and therefore $\int_{\mathbf{r}(D)} f dS = \int_{\mathbf{r}_1(D_1)} f dS$ as desired.

□

Changing axis orientation:

In most cases it is easier to do surface integrals (particularly flux integrals) using the formula where we describe a surface S as a graph of a function $z = f(x, y)$. For flux integrals oriented up, this is:

$$\int\int_{S_1} \mathbf{F}\cdot d\mathbf{S} = \int\int_D -Pg_x - Qg_y + RdA$$

This is because parametrizing a surface in the first place takes time. Then taking the cross $\mathbf{r}_u \times \mathbf{r}_v$ takes time. Even after doing those two steps the integrand after taking $\mathbf{F}\cdot\mathbf{r}_u \times \mathbf{r}_v$ is often worse than the integral in the formula above unless your parametrization was carefully chosen (in which case the integrand probably is better for the parametrized form).

However, it is frequently true that a surface is not easily described as the graph of a function $z = f(x, y)$ but it can be described as $y = f(x, z)$ or $x = f(y, z)$. In such cases the formula above can still be used with minor changes. As discussed in earlier sections, we can often just switch the axes and switch the corresponding variables in integrands and set up bounds that way so that our picture is more convenient. When working with vector fields, however, we have to make sure that the components of those vector fields are consistent with the variable switch. If we were to just switch the axes and switch the variables then we would still get an equivalent integrand with a nicer picture, but for the vector field $\mathbf{F} =< P(x, y, z), Q(x, y, z), R(x, y, z) >$, if we switch which variable is x and which is z, for instance, then as long as we switch the region accordingly we will have the same values for the entries of $\mathbf{F}$ (at points where coordinates are exchanged), but the direction $\mathbf{F}$ points will not be the same relative to the new position of the axes. This is because $\mathbf{F}$ still says that P is pointing in the x direction, which has, in the new axes, taken the place of the z-direction. So, in addition to switching x with z we must also switch the coordinates P and R of $\mathbf{F}$ so that the new $\mathbf{F} =< R(z, y, x), Q(z, y, x), P(z, y, x) >$ corresponds to the newly positioned axes.

The formulas are as follows in the case of the flux integral above:

If S_1 is the graph of $y = g(x, z)$ over D in the xz-plane then the flux integral of $\mathbf{F} =< P, Q, R >$ through S_1 oriented in the positive y direction is:

$$\int\int_{S_1} \mathbf{F}\cdot d\mathbf{S} = \int\int_D -Pg_x - Rg_z + QdA$$

If S_1 is the graph of $x = g(y, z)$ over D in the yz-plane then the flux integral of $\mathbf{F} =< P, Q, R >$ through S_1 oriented in the positive x direction is:

$$\int\int_{S_1} \mathbf{F}\cdot d\mathbf{S} = \int\int_D -Qg_y - Rg_z + PdA$$

There are similar corresponding formulas for surface area and surface integrals, though these depend less on a vector field so they may be more apparent:

If S_1 is the graph of $y = g(x, z)$ over D in the xz-plane then the surface area of S_1 is:

$$A = \int\int_D \sqrt{1 + g_x^2 + g_z^2} dA$$

The corresponding surface integral formula is:

$$\int\int_{S_1} f dS = \int\int_D f(x, g(x, z), z)\sqrt{1 + g_x^2 + g_z^2} dA$$

If S_1 is the graph of $x = g(y, z)$ over D in the yz-plane then the surface area of S_1 is:

$$A = \int\int_D \sqrt{1 + g_y^2 + g_z^2} dA$$

The corresponding surface integral formula is:

$$\int\int_{S_1} f dS = \int\int_D f(g(y, z), y, z)\sqrt{1 + g_y^2 + g_z^2} dA$$

Stokes's Theorem

Stokes's Theorem gives us a way to use a flux integral of the curl of a vector field through a surface to determine the line integral of a vector field along a closed path which is the boundary for that surface. Since curl vectors are often much simpler than the original vector field, this is helpful. This also lets us prove results which have been stated earlier relating circulation at a point to the curl vector. Here is the statement of the result. We will only prove the result for special cases. To prove the general form of the theorem would require significant build up and probably be almost as long as the Change of Variables Theorem, and it probably isn't worth it for purposes of a text that is primarily focused on calculus methods because for just about every surface most readers would want to use Stokes's Theorem for the special case we will prove is sufficient. In other words, the increase in the generality of the result does not seem to be worth the increase in the complication of the argument and the higher likelihood that a student would become lost in this case. The generalized form of Stokes's Theorem is even more complicated than the one listed here as the general form of Stokes's Theorem, but it is exceptionally useful in certain areas of mathematics and readers who are interested are encouraged to study it.

Theorem 13.21. *Stokes's Theorem. Let $\boldsymbol{F} =< P, Q, R >$ be a C^1 vector field on a connected open set D containing a piecewise smooth closed surface S_1 oriented upwards and its boundary (with respect to the subspace topology) is a smooth closed curve C which is oriented counterclockwise as viewed from above. Then $\int\int_{S_1} curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}$.*

There are some issues with the statement of this theorem. The first is that we have not defined piecewise smooth surface. The second is that it isn't all that clear what we mean by "with respect to the subspace topology." A piecewise smooth surface is a union of smooth surfaces and closures of smooth surfaces that only intersect along their boundaries, but this is a more general construction than we wish to work with in this text. Recall that for a smooth surface, if we restrict ourselves to looking at a portion of the surface in a small enough ball then the surface is locally the graph of a C^1 function (either $z = f(x, y)$ or $x = g(y, z)$ or $y = h(x, z)$ over some Jordan region bounded by a smooth curve gives the same points as the intersection of the surface with the ball). As a result, we would expect that most surfaces we are likely to encounter can be obtained by taking surfaces which are graphs of functions over connected Jordan regions which are closures of open sets and joining them together along their boundaries. This motivates us to define the following:

Definition 117

Let E be a connected open Jordan region in $\mathbb{R}^2$. Let $D = \overline{U} \subset E$, where U is open and connected and D is a Jordan region whose boundary is a piecewise smooth closed curve. Let f be a C^1 function on E so that $\mathbf{r}_z(u, v) = (u, v, f(u, v))$ is a parametrization for the regular surface $\mathbf{r}(E)$. Then $\mathbf{r}_z(D)$ is a *standard smooth graph surface* in the variable z over the region D in the xy-plane. We also say that $\mathbf{r}_x(u, v) = (f(u, v), u, v)$ and $\mathbf{r}_y(u, v) = (u, f(u, v), v)$ are standard smooth graph surfaces in the variables x and y respectively over the region D in the yz and xz planes respectively. We say that a point $\mathbf{p} \in \mathbf{r}(E)$ is in the boundary of a set $S \subset \mathbf{r}(E)$ *in the subspace topology* on $\mathbf{r}(E)$ if, for every $\epsilon > 0$, the ball $B_\epsilon(\mathbf{p})$ contains a point in S and a point of $\mathbf{r}(E)$ which is not contained in S. To avoid repetition of "with respect to the subspace topology on $\mathbf{r}(E)$" we will refer to the boundary of S with respect to the subspace topology on $\mathbf{r}(E)$ as the *manifold boundary* of S. We refer to the maps $\mathbf{r}_x$ or $\mathbf{r}_y$ or $\mathbf{r}_z$ as *standard graph functions* with respect to variables x, y and z respectively.

If $W = \bigcup_{i=1}^{m} \mathbf{r}_i(D_i)$, for standard graph functions $\mathbf{r}_i$ on open sets E_i containing D_i for each i, is a union of standard smooth graph surfaces, then we say that a point $\mathbf{p}$ is in the boundary of W with respect to the subspace topology on $\bigcup_{i=1}^{m} \mathbf{r})i(E_i)$ (or just in the manifold boundary) if every open ball containing $\mathbf{p}$ contains a point of W and a point of $\bigcup_{i=1}^{m} \mathbf{r}_i(E_i)$ which is not in W.

It may be asked why we have to talk about manifold boundary. Well, we are looking at the boundary within the surface itself. The topological boundary (the boundary we have been discussing in all the other sections of the text) of a regular surface in $\mathbb{R}^3$ is the entire surface in every case. We want to specifically look at the edge of the surface, which is the boundary with respect to an extension of a surface to a slightly larger surface within the surface itself. In this context we could talk about the manifold boundary of the portion of a paraboloid $z = 1 - x^2 - y^2$ for $z \geq 0$ being the unit circle on the xy-plane where the paraboloid was cut off (but every open ball in $\mathbb{R}^3$ containing a point on that section of a paraboloid will contain points on the paraboloid and points of space that are not on the paraboloid, which would would mean the topological boundary (in three space) of the surface would be the entire surface).

When we refer to a manifold boundary for a surface that could be represeneted as a piecewise smooth graph surface where a point would be in the boundary then the boundary of the surface with the unspecified parametrization includes that point. Thus, we would say that the manifold boundary of the surface S consisting of the portion of the paraboloid $z = 1 - x^2 - y^2$ where $z \geq 0$ was the circle C defined by $x^2 + y^2 = 1$ in the xy-plane because we could parametrize the surface as $\mathbf{r}(u, v, u^2 + v^2)$ over $u^2 + v^2 < 100$, and with respect to that parametrization the boundary of S would be C. Essentially, we extend the domain of the parametrization to an open set when we take the manifold boundary.

If $\mathbf{l}$ is the smooth closed curve C which is the boundary of D in the above definition then $\mathbf{r}_z(C)$ is the manifold boundary of $\mathbf{r}_z(D)$. If C is oriented counterclockwise then we say that $\mathbf{r} \circ \mathbf{l}$ or oriented counterclockwise as viewed from above (or if we replace z by x or y then as viewed from the positive x or positive y direction).

Theorem 13.22. *Let E be a connected open Jordan region in $\mathbb{R}^2$. Let $D = \overline{U} \subset E$, where U is open and connected and D is a Jordan region whose boundary is a counter clockwise oriented piecewise smooth closed curve C parametrized by $\boldsymbol{l} : [a, b] \to \mathbb{R}^2$. Let f be a C^1 function on E so that $\boldsymbol{r}_z(u, v) = (u, v, f(u, v))$ is a standard graph parametrization for the regular surface $\boldsymbol{r}_z(E)$. Then the manifold boundary of $S = \boldsymbol{r}_z(D)$ is $\boldsymbol{r}_z(C)$ which is parametrized by $\boldsymbol{r}_z \circ \boldsymbol{l} : [a, b] \to \mathbb{R}^3$.*

Likewise, if $\boldsymbol{r}_y(u, v) = (u, f(u, v), v)$ or $\boldsymbol{r}_x(u, v) = (f(u, v), u, v)$ are the standard graph parametrizations for the surface then $\boldsymbol{r}_y(C)$ or $\boldsymbol{r}_x(C)$ are the manifold boundaries for the $\boldsymbol{r}_y(D)$ or $\boldsymbol{r}_x(D)$ respectively.

Proof. We will only prove this for the first case. The other two cases are the same up to a re-labeling for variables. First, observe that the path $\mathbf{r}_z \circ \mathbf{l}$ is piecewise smooth since whenever $\sqrt{x'(t)^2 + y'(t)^2} > 0$ it is also true that $\sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2} > 0$ (where $z(t) = f(x(t), y(t))$) and the composition of continuously differentiable functions is continuously differentiable. Let $\mathbf{p} = (x, y, f(x, y)) \in \mathbf{r}_z(C)$. Then $(x, y) \in C$. Let $\epsilon > 0$ be small enough so that $B_\epsilon((x, y)) \subset E$ in $\mathbb{R}^2$. Since f is continuous we can choose $\delta < \dfrac{\epsilon}{2}$ so that if $|(x, y) - (s, t)| < \delta$ then $|f(x, y) - f(s, t)| < \dfrac{\epsilon}{2}$. Then $B_\delta((x, y))$ contains a point $(s_1, t_1) \notin D$ and a point $(s_2, t_2) \in D$, which means that $(s_1, t_1, f(s_1, t_1)) \in B_\epsilon(\mathbf{p}) \setminus \mathbf{r}_z(D)$ and $(s_2, t_2, f(s_2, t_2)) \in B_\epsilon(\mathbf{p}) \cap \mathbf{r}_z(D)$. Thus, $\mathbf{r}_z(C)$ is contained in the manifold boundary of $\mathbf{r}_z(D)$.

If a point $(x, y, z) \in \mathbf{r}_z(E) \setminus \mathbf{r}_z(C)$ then $(x, y) \notin C$, which means that there is some $\gamma > 0$ so that either $B_\gamma((x, y)) \cap D = \emptyset$, in which case $B_\gamma(x, y, z) \cap \mathbf{r}_z(D) = \emptyset$, or $B_\gamma((x, y)) \subset D$, in which case $B_\gamma((x, y, z)) \cap \mathbf{r}_z(E) \subset \mathbf{r}_z(D)$. Thus, the manifold boundary of $\mathbf{r}_z(D)$ is $\mathbf{r}_z(C)$. □

It is immediate from the definition that standard smooth graph surfaces are standard surfaces. Standard smooth graph surfaces have a boundary (in the subspace topology) that is easy to see. It is just the image of the boundary of D under $\mathbf{r}_z$ (or $\mathbf{r}_x$ or $\mathbf{r}_y$, depending on which variable the surfaces is a smooth graph in).

Definition 118

Let E be a connected open Jordan region in $\mathbb{R}^2$. Let $D = \overline{U} \subset E$, where U is open and connected and D is a Jordan region whose boundary is a counter clockwise oriented piecewise smooth closed curve C parametrized by $\mathbf{l} : [a,b] \to \mathbb{R}^2$. Let f be a C^1 function on E so that $\mathbf{r}_z(u,v) = (u, v, f(u,v))$ is a standard graph parametrization for the regular surface $\mathbf{r}_z(E)$. Then we say that the piecewise simple closed curve Q defined by $\mathbf{r}_z \circ \mathbf{l} : [a,b] \to \mathbb{R}^3$ (whose trace is the manifold boundary) is *oriented counterclockwise as viewed from above.* If we replace $\mathbf{r}_z$ with $\mathbf{r}_y(u,v) = (u, f(u,v), v)$ or $\mathbf{r}_x(u,v) = (f(u,v), u, v)$ in this definition then we say that Q is oriented counterclockwise as viewed from the positive y or x direction respectively.

We define a piecewise smooth graph surface inductively. First, a standard smooth graph surface is a piecewise smooth graph surface. Next, assume that the union of smooth graph surfaces $S = S_1 \cup S_2 \cup \ldots \cup S_{k-1}$ is a piecewise smooth graph surface whose manifold boundary is a piecewise smooth closed curve C_{k-1} with parametrization $\mathbf{v}$. Let S_k be a smooth graph surface with manifold boundary C_k with parametrization $\mathbf{w}$ so that $S_k \cap S \subseteq C_k \cap C_{k-1} = Q$ with parametrization $\mathbf{q} : [c,d] \to \mathbb{R}^3$, a smooth curve whose orientation induced by $\mathbf{w}$ and $\mathbf{v}$ are opposite to each other, and $S_k \cup S$ has manifold boundary $C = (C_{k-1} \cup C_k \setminus \mathbf{q}((c,d))$. Further assume that C has a parametrization $\mathbf{r}$ so that the orientation of $C_{k-1} \setminus \mathbf{q}((c,d))$ induced by $\mathbf{r}$ is the same as the orientation induced by $\mathbf{v}$ and the orientation of $C_k \setminus \mathbf{q}((c,d))$ induced by $\mathbf{r}$ is the same as the orientation induced by $\mathbf{w}$. Then $S \cup S_k$ is a piecewise smooth graph surface.

We will now prove Stokes's theorem for a standard smooth graph surface. It is convenient to have a notation for manifold boundary, so we will use $\partial_M(S)$ to denote the boundary of S with respect to subspace topology on the set M.

Theorem 13.23. *Special Case of Stokes's Theorem for standard smooth graph surfaces. Let E be a connected open Jordan region in $\mathbb{R}^2$. Let $D = \overline{O} \subset E$, where O is open and connected and D is a Jordan region whose boundary is a counter clockwise oriented piecewise smooth closed curve L parametrized by $\boldsymbol{l}(t) =< x(t), y(t) >: [a,b] \to \mathbb{R}^2$. Let g be a C^1 function on E so that $\boldsymbol{r}(u,v) = (u, v, g(u,v))$ is a standard graph parametrization for the regular surface $S_1 = \boldsymbol{r}(E)$ with respect to variable z. Let C be the counter clockwise as viewed from above piecewise smooth curve parametrized by $\boldsymbol{r} \circ \boldsymbol{l}(t) =< x(t), y(t), g((x(t), y(t)) >: [a,b] \to \mathbb{R}^3$. Let $F =< P, Q, R >$ be a C^1 vector field on an open set U containing S_1.*

Then $\displaystyle\int\int_{S_1} curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}.$

Likewise, if S_1 is $\boldsymbol{r}_y(D)$ or $\boldsymbol{r}_x(D)$, where $\boldsymbol{r}_y(u,v) = (u, f(u,v), v)$ or $\boldsymbol{r}_x(u,v) = (f(u,v), u, v)$ are the standard graph parametrizations for the surface then if the orientation of the manifold boundary curve is counterlockwise as viewed from the positive x or y direction respectively then $\displaystyle\int\int_{S_1} curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}.$

Proof. The argument is the same in each direction up to a change in coordinate labeling, so we only prove this result for the first stated case.

We define $P^{(xy)}(x,y) = P(x,y,g(x,y))$, $Q^{(xy)}(x,y) = Q(x,y,g(x,y))$ and $R^{(xy)}(x,y) = R(x,y,g(x,y))$ on D and observe that $P(\mathbf{r}(t)) = P^{(xy)}(\mathbf{l}(t))$, $Q(\mathbf{r}(t)) = Q^{(xy)}(\mathbf{l}(t))$, and $R(\mathbf{r}(t)) = R^{(xy)}(\mathbf{l}(t))$ for all $t \in [a,b]$. We have $\int_C \mathbf{F}\cdot d\mathbf{r} = \int_a^b P(\mathbf{r}(t))x'(t)dt + \int_a^b Q(\mathbf{r}(t))y'(t)dt + \int_a^b R(\mathbf{r}(t))z'(t)dt$. Since $z = g(x,y)$ we can use the Chain Rule to write $z'(t) = \frac{\partial g}{\partial x}\frac{dx}{dt} + \frac{\partial g}{\partial y}\frac{dy}{dt}$. Hence, we can write the path integral as $\int_a^b (P(\mathbf{r}(t)) + g_xR(\mathbf{r}(t)))x'(t) + (Q(\mathbf{r}(t)) + g_yR(\mathbf{r}(t)))y'(t)dt = \int_a^b (P^{(xy)}(\mathbf{l}(t)) + g_xR^{(xy)}(\mathbf{l}(t)))x'(t) + (Q^{(xy)}(\mathbf{l}(t)) + g_yR^{(xy)}(\mathbf{l}(t)))y'(t)dt$.

So, by Green's Theorem, we can write $\int_a^b (P(\mathbf{r}(t))+g_xR(\mathbf{r}(t)))x'(t)+(Q(\mathbf{r}(t))+g_yR(\mathbf{r}(t)))y'(t)dt$ $= \int_L (P^{(xy)}+R^{(xy)}g_x)dx+(Q^{(xy)}+R^{(xy)}g_y)dy = \int\int_D \frac{\partial(Q^{(xy)}+R^{(xy)}g_y)}{\partial x} - \frac{\partial(P^{(xy)}+R^{(xy)}g_x)}{\partial y}dA$ $= \int\int_D Q_x^{(xy)}+R_x^{(xy)}g_y+R^{(xy)}g_{yx}-P_y^{(xy)}-R_y^{(xy)}g_x-R^{(xy)}g_{xy}dA$. By Clairaut's Theorem this simplifies to $= \int\int_D Q_x^{(xy)} - P_y^{(xy)} + R_x^{(xy)}g_y - R_y^{(xy)}g_xdA$. Also, if $h(x,y,z) = (x,y,g(x,y))$ then $P^{(xy)}(x,y) = (P \circ h)(x,y)$, $Q^{(xy)}(x,y) = (Q \circ h)(x,y)$ and $R^{(xy)}(x,y) = (R \circ h)(x,y)$. Hence, by the Chain Rule we have that $Q_x^{(xy)}((x,y)) = \nabla Q(h((x,y)))\cdot h_x((x,y)) = Q_x(x,y,g(x,y)) + Q_z(x,y,g(x,y))g_x(x,y)$. Similarly, $P_y^{(xy)}((x,y)) = P_y(x,y,g(x,y))+P_z(x,y,g(x,y))g_y(x,y)$, $R_x^{(xy)}((x,y)) = R_x(x,y,g(x,y))+R_z(x,y,g(x,y))g_x(x,y)$ and $R_y^{(xy)}((x,y)) = R_y(x,y,g(x,y)) + R_z(x,y,g(x,y))g_y(x,y)$. Hence, this integral can be written as $\int_D Q_x^{(xy)} - P_y^{(xy)} + R_x^{(xy)}g_y - R_y^{(xy)}g_xdA = \int\int_D Q_x+Q_zg_x-P_y-P_zg_y+R_xg_y+R_zg_xg_y-R_yg_x-R_zg_yg_xdA = \int\int_D -(R_y - Q_z)g_x - (P_z - R_x)g_y + Q_x - P_ydA = \int\int_{S_1} \text{curl}(\mathbf{F}) \cdot d\mathbf{S}$.

□

Example 13.21. *Evaluate the path integral* $\int_C \boldsymbol{F}\cdot \boldsymbol{dr}$ *where* C *is the rectangle with vertices* $(0,0,2)$*,* $(2,0,2)$*,* $(2,2,2)$ *and* $(0,2,2)$*, starting at* $(0,0,2)$ *and traversed clockwise as viewed from above, and*

$$\boldsymbol{F}(x,y,z) =< z + e^{x^2}, 5x + y\cos(y^2), z^2 + 4x >$$

Solution. The path is a closed curve in three dimensions, and the curl of the vector field is simple, so we use Stokes's Theorem. The simplest surface S_1 that this curve bounds appears to be the surface of the graph of $g(x,y) = 2$ over $D = [0,2] \times [0,2]$. Since the curve is traversed clockwise as viewed from above, Stokes's Theorem would give us that $\int_C \mathbf{F}\cdot d\mathbf{r} = -\iint_{S_1} \text{curl}(\mathbf{F}) \cdot d\mathbf{S}$, where S_1 is oriented upwards, which is equal to $\iint_D Pg_x + Qg_y - RdA$, where $< P, Q, R >$ represent the coordinates of the curl of $\mathbf{F}$, which is:

$$\text{curl}(\mathbf{F}) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ z + e^{x^2} & 5x + y\cos(y^2) & z^2 + 4x \end{bmatrix} =< 0, -3, 5 >.$$ Thus, the flux integral is

equal to $\iint_D 0(0) + -3(0) - 5dA = -5(4) = -20.$

□

We would like to extend Stokes's Theorem to a fairly general case which will encompass most surfaces we might be interested in, so we include the following theorem.

Theorem 13.24. *Let S_1 be a piecewise smooth graph surface with piecewise smooth closed curve manifold boundary C given by $\boldsymbol{v} : [a_1, b_1] \to \mathbb{R}^3$ so that for every C^1 vector field F on an open set containing S_1 it is true that $\int\int_{S_1} curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_C \boldsymbol{F} \cdot d\boldsymbol{r}$. Let M be a standard smooth graph surface with manifold boundary B parametrized by the piecewise smooth curve $\boldsymbol{w} : [a_2, b_2]$ so that $M \cap S_1$ is a simple piecewise smooth curve I given by $\boldsymbol{l} : [a_3, b_3] \to \mathbb{R}^3$, where the orientation given by $\boldsymbol{l}$ is the same as the orientation induced by $\boldsymbol{w}$ and the opposite of the orientation induced by $\boldsymbol{v}$, and the manifold boundary of $S = S_1 \cup M$ is $K = (C \cup B) \setminus \boldsymbol{l}((a_3, b_3))$ which has a piecewise smooth parametrization $\boldsymbol{r} : [a, b] \to \mathbb{R}^3$ so that the orientation of $C \setminus \boldsymbol{l}((a_3, b_3))$ induced by $\boldsymbol{r}$ is the same as the orientation induced by $\boldsymbol{v}$ and the orientation of $B \setminus I$ induced by $\boldsymbol{r}$ is the same as the orientation induced by $\boldsymbol{w}$. Then for every C^1 vector field F on an open set containing S it is true that $\int\int_S curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_K \boldsymbol{F} \cdot d\boldsymbol{r}$.*

Moreover, if S is any piecewise smooth graph surface then S has a manifold boundary curve K so that $\int\int_S curl(\boldsymbol{F}) \cdot d\boldsymbol{S} = \int_K \boldsymbol{F} \cdot d\boldsymbol{r}$.

Proof. We know that $\int_K \mathbf{F} \cdot d\mathbf{r} = \int_{C\setminus\mathbf{l}((a_3,b_3))} \mathbf{F} \cdot d\mathbf{r} + \int_{B\setminus\mathbf{l}((a_3,b_3))} \mathbf{F} \cdot d\mathbf{r} = \int_{C\setminus\mathbf{l}((a_3,b_3))} \mathbf{F} \cdot d\mathbf{r} + \int_{B\setminus\mathbf{l}((a_3,b_3))} \mathbf{F} \cdot d\mathbf{r} + \int_I \mathbf{F} \cdot d\mathbf{r} - \int_I \mathbf{F} \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot d\mathbf{r} + \int_B \mathbf{F} \cdot d\mathbf{r} = \int\int_{S_1} \text{curl}(\mathbf{F}) \cdot d\mathbf{S} + \int\int_M \text{curl}(\mathbf{F}) \cdot d\mathbf{S} = \int\int_S \text{curl}(\mathbf{F}) \cdot d\mathbf{S}$.

We know Stokes's Theorem is true for a standard smooth graph surface. Proceeding inductively, by definition we see that any piecewise smooth graph surface is a union as described by definition, which means, inductively, that if we assume $\int\int_S \text{curl}(\mathbf{F}) \cdot d\mathbf{S} = \int_K \mathbf{F} \cdot d\mathbf{r}$ when S is a piecewise smooth graph surface which is a union of $k-1$ standard smooth graph surfaces then it follows that is S is the union of k standard graph surfaces that $\int\int_S \text{curl}(\mathbf{F}) \cdot d\mathbf{S} = \int_K \mathbf{F} \cdot d\mathbf{r}$ as well. We conclude that $\int\int_S \text{curl}(\mathbf{F}) \cdot d\mathbf{S} = \int_K \mathbf{F} \cdot d\mathbf{r}$ for any piecewise smooth graph surface (for one of the two possible orientations of K). □

From the preceding theorem, we see that if we can construct a piecewise smooth graph surface by adding in a new standard smooth surface repeatedly so that the intersection of the previous piecewise smooth graph surface with the new surface is a simple piecewise smooth curve with opposite induced orientation along the intersection curve, then the union will be a new piecewise smooth surface to which Stokes's Theorem applies. Since a standard surface which is oriented clockwise from above and a standard surface oriented clockwise from the

x or y direction which intersect along a piecewise smooth simple curve will always have opposite orientations along the intersection we can just keep appending standard smooth graph surfaces to construct more complex surfaces to which Stokes's Theorem applies.

It is helpful to be able to use curl to decide whether a vector field is conservative. Before we can prove this, we must check that we can parametrize a triangular disk.

Theorem 13.25. *Let T be the triangle with vertices $(0,0,0)$, (x_1, y_1, z_1) and (x_2, y_2, z_2). Then there is a C^1 parametrization $\boldsymbol{r}$ on a Jordan region D so that $\boldsymbol{r}(D)$ is the triangular disk bounded by T, and $\displaystyle\int_C \boldsymbol{F} \cdot \boldsymbol{dr} = 0$ if curl$(\boldsymbol{F}) = \boldsymbol{0}$.*

Solution. Let $\mathbf{r}(u, v) = u < x_1, y_1, z_1 > +v < x_2, y_2, z_2 >$ with $0 \le u \le 1$ and $0 \le v \le 1-u$. This parametrization includes all points on the line segment from $(0,0,0)$ to (x_1, y_1, z_1) when $v = 0$. When $u = 0$ we have all points on the line segment from $(0,0,0)to(x_2, y_2, z_2)$. Any point on the line segment from (x_1, y_1, z_1) to (x_2, y_2, z_2) is of the form $(x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_2), z_1 + t(z_2 - z_1) >= (1-t) < x_1, y_1, z_1 > +t < x_2, y_2, z_2 >$ for some $0 \le t \le 1$. If we set $u = t$ then when $v = 1-u$ this is $v < x_1, y_1, z_1 > +u < x_2, y_2, z_2 >$, so all three edges of the triangle are in the parametrization. Since the largest value of v for a given u is $1-u$, we include no points in the parametrization of the form $u < x_2, y_2, z_2 > +k < x_1, y_1, z_1 >$ where $k > 1-u$, which means that we only add vector lengths to each point along the line segment from $(0,0,0)$ to (x_2, y_2, z_2) which give points in the triangle (no points beyond the triangle, but all the points within the triangle of that form). Every point in the triangle is in the parallelogram with u and v as edges, which is $\mathbf{p}(u, v) = u < x_1, y_1, z_1 > +v < x_2, y_2, z_2 >$ with $0 \le u \le 1$ and $0 \le v \le 1$. Hence, the parametrization includes exactly the points of this parallelogram of form $u < x_2, y_2, z_2 > +k < x_1, y_1, z_1 >$ which are in the triangle, so $\mathbf{r}$ parametrizes the entire triangular disk.

If curl$(\mathbf{F}) = \mathbf{0}$ then $\displaystyle\int_C \mathbf{F} \cdot \mathbf{dr} = 0$ by Stokes's Theorem, since $\mathbf{C}$ is a piecewise-smooth path.

□

Theorem 13.26. *(a) Let $\boldsymbol{F} =< P, Q, R >$ be a C^1 vector field on $\mathbb{R}^3$. Then $\boldsymbol{F}$ is conservative if and only if curl$(\boldsymbol{F}) = \boldsymbol{0}$.*

(b) Let $\boldsymbol{G} =< P, Q >$ be a C^1 vector field on $\mathbb{R}^2$. Then $\boldsymbol{G}$ is conservative if and only if curl$(\boldsymbol{F}) = \boldsymbol{0}$.

Proof. (a) First, assume that $\mathbf{F}$ is conservative. Since $\mathbf{F}$ is conservative, there is a function f so that $\nabla f = \mathbf{F} =< f_x, f_y, f_z >$. Hence, curl$(\mathbf{F}) =< f_{zy} - f_{yz}, f_{xz} - f_{zx}, f_{yx} - f_{xy} >=< 0, 0, 0 >$ by Clairaut's Theorem.

To prove the other direction, we first observe that if T is a closed path which is a triangle then T bounds a triangular disk D which is always a standard smooth graph surface by Theorem 13.25), so by our version of Stokes's Theorem we know that if C is the closed path along the triangle T then $\displaystyle\int_T \mathbf{F} \cdot \mathbf{dr} = \int\int_D \text{curl}(\mathbf{F}) \cdot \mathbf{dS} = 0$ since curl$(\mathbf{F}) = 0$. For a given point (x, y, z), for any point (a, b, c) which is not colinear with (x, y, z) and the origin we know that if C is the path along the triangle whose vertices are the origin, (a, b, c) and

(x, y, z) then $\displaystyle\int_C \mathbf{F}\cdot\mathbf{dr} = 0$, which means that the integral along the path C_1 from the origin to (x, y, z) is the same as the integral along the path C_2 from the origin to (a, b, c) followed by the path C_3 from the point (a, b, c) to the point (x, y, z). The requirement that the points not be colinear is only for applying Stokes's theorem (if they are colinear then the line integral along C is still zero since we know that the sum of the integrals along a line in one direction and then returning to its point of origin still gives an integral of zero). As long as a path consists of only two line segments, then the integral is independent of the choice of such paths. We define $f(x, y, z) = \displaystyle\int_{L(x,y,z)} \mathbf{F}\cdot\mathbf{dr}$, where $L(x, y, z)$ is the line segment from the origin to (x, y, z). Let L be the line segment from $(0, 0, 0)$ to $(x - 1, y, z)$. Let L_u be the line segment from $\mathbf{r}(t) =< x - 1 + t, y, z >$, over $0 \le t \le u$, and note that L_1 is the line segment from $(x - 1, y, z)$ to (x, y, z). Then $f_x(x, y, z) \displaystyle\lim_{u\to 0} \frac{f(x + u, y, z) - f(x, y, z)}{u} =$
$\displaystyle\lim_{u\to 0} \frac{L + \int_{L_u} \mathbf{F}\cdot\mathbf{dr} - (L + \int_{L_1} \mathbf{F}\cdot\mathbf{dr})}{u} = \lim_{u\to 0} \frac{\int_1^u \mathbf{F}(x - 1 + t, y, z)\cdot < 1, 0, 0 > dt}{u}$
$= \displaystyle\lim_{u\to 0} \frac{\int_1^u P(x - 1 + t, y, z)dt}{u} = \frac{\partial}{\partial u}\int_1^u P(x - 1 + t, y, z)dt = P(x - 1 + u, y, z)$. Hence, the derivative of f with respect to x when $u = 1$ is $f_x(x, y, z) = P(x, y, z)$ by the Fundamental Theorem of Calculus. By renaming variables and repeating the argument we see that $f_y(x, y, z) = Q(x, y, z)$ and $f_z(x, y, z) = R(x, y, z)$. Hence, $\triangledown f = \mathbf{F}$.

(b) First, assume that $\mathbf{F}$ is conservative. Since $\mathbf{F}$ is conservative, there is a function f so that $\triangledown f = \mathbf{F} =< f_x, f_y >$. Hence, curl$(\mathbf{F}) = f_{yx} - f_{xy} = 0$ by Clairaut's Theorem.

Using the argument above, if we define $\mathbf{F}(x, y, z) =< P, Q, 0 >$ (where P and Q are functions of only x and y just as in the definition of G) then curl$(\mathbf{F}) =< 0, 0, Q_x - P_y >= 0$. Hence, F is path inedpendent. Let C_1 and C_2 be piecewise smooth paths parametrized by $\mathbf{r}_1 : [a_1, b_1] \to \mathbb{R}^2$ and $\mathbf{r}_2 : [a_2, b_2] \to \mathbb{R}^2$ respectively, and let K_1 and K_2 be corresponding paths parametrized by $\mathbf{R}_1 : [a_1, b_1] \to \mathbb{R}^3$ defined by $\mathbf{R}_1(t) =< \mathbf{r}_1(t), 0 >$ and $\mathbf{R}_2 : [a_2, b_2] \to \mathbb{R}^3$ defined by $\mathbf{R}_2(t) =< \mathbf{r}_2(t), 0 >$ repectively, from $(x_1, y_1, 0)$ to $(x_2, y_2, 0)$. Let $\mathbf{r}_1(t) =< x_1(t), y_1(t) >$ and let $\mathbf{r}_2(t) =< x_2(t), y_2(t) >$. Then $\displaystyle\int_{K_1} \mathbf{F}\cdot\mathbf{dr} = \int_{a_1}^{b_1} < P, Q, 0 > \cdot < x_1'(t), y_1'(t), 0 > dt = \int_{a_1}^{b_1} < P, Q > \cdot < x_1'(t), y_1'(t) > dt = \int_{C_1} \mathbf{G}\cdot\mathbf{dr}$, and $\displaystyle\int_{K_2} \mathbf{F}\cdot\mathbf{dr} = \int_{a_2}^{b_2} < P, Q, 0 > \cdot < x_2'(t), y_2'(t), 0 > dt = \int_{a_2}^{b_2} < P, Q > \cdot < x_2'(t), y_2'(t) > dt = \int_{C_2} \mathbf{G}\cdot\mathbf{dr}$.
Since we know that $\displaystyle\int_{K_1} \mathbf{F}\cdot\mathbf{dr} = \int_{K_2} \mathbf{F}\cdot\mathbf{dr}$, it must follow that $\displaystyle\int_{C_1} \mathbf{G}\cdot\mathbf{dr} = \int_{C_2} \mathbf{G}\cdot\mathbf{dr}$, so $\mathbf{G}$ is path independent and therefore conservative.

□

Gauss's Theorem

The Divergence Theorem, also known as Gauss's Theorem, essentially says that if you add up the divergence per unit volume within a solid then the sum of the divergence times volumes will add up to the net flex out through the surface bounding the solid. Thus, if you think of it in terms of the flow of a gas, the net expansion is the flow of the gas out through the boundary. The more general statement of the Divergence Theorem is as follows:

Theorem 13.27. *Divergence Theorem. Let $E \subseteq \mathbb{R}^3$, where E is a compact set with piecewise smooth boundary surface S_1. If there is an open set U containing E on which vector field $F =< P, Q, R >$ is C^1 then $\int\int\int_E div(\boldsymbol{F})dV = \int\int_{S_1} \boldsymbol{F} \cdot d\boldsymbol{S}$, where S_1 is oriented outwards.*

Since our focus in this course is on simpler regions, we will prove the following special case of the Divergence Theorem.

Theorem 13.28. *Special Case of the Divergence Theorem. Let $E \subseteq \mathbb{R}^3$ by a region of type 1, 2 and 3 whose boundary functions are C^1, where S is the boundary of E. If there is an open set U containing E on which vector field $F =< P, Q, R >$ is C^1 then $\int\int\int_E div(\boldsymbol{F})dV = \int\int_S \boldsymbol{F} \cdot d\boldsymbol{S}$, where S is oriented outwards.*

Proof. Since E is of types one, two and three with smooth boundary functions, it follows that there are C^1 functions $z = h_1(x, y)$, $z = h_2(x, y)$ with E equal to the solid between these two functions over a domain D_z in the xy plane which is a region of types one and two (so $E = \{(x, y, z) \in \mathbb{R}^3 | (x, y) \in D$ and $h_1(x, y) \leq z \leq h_2(x, y)\}$, and likewise there are C^1 functions $y = g_1(x, z)$, $y = g_2(x, z)$ with E equal to the solid between these two functions over a domain D_y in the xz plane which is a region of types one and two, and also there are C^1 functions $x = f_1(y, z)$, $x = f_2(y, z)$ with E equal to the solid between these two functions over a domain D_x in the yz plane which is a region of types one and two.

We focus on the first description for now (E is equal to the solid between $z = h_1(x, y)$, $z = h_2(x, y)$ over domain D_z). Then the boundary S of E is a union of three surfaces: $S_1 = \{(x, y, h_1(x, y)) \in \mathbb{R}^3 | (x, y) \in D_z\}$ oriented downwards, $S_2 = \{(x, y, h_2(x, y)) \in \mathbb{R}^3 | (x, y) \in D_z\}$ oriented upwards, and $S_3 = \{(x, y, z) | (x, y) \in \partial(D)$ and $h_1(x, y) \leq z \leq h_2(x, y)\}$ oriented outwards (away from D). The argument that this is the boundary is the same as arguments we have already given in the proof that a type three region is a Jordan region.

We know that $\int\int\int_E \text{div}(\mathbf{F})dV = \int\int\int_E P_x dV + \int\int\int_E Q_y dV + \int\int\int_E R_z dV$. Looking at the third of these integrals we see that $\int\int\int_E R_z dV = \int\int_{D_z} \int_{h_1(x,y)}^{h_2(x,y)} R_z dz dA$ $= \int\int_{D_z} R(h_2(x, y)) - R(h_1(x, y))dA$ by the Fundamental Theorem of Calculus. The flux integral $\int\int_S \mathbf{F} \cdot d\mathbf{S} = \int\int_S (P\mathbf{i} + Q\mathbf{j} + R\mathbf{k}) \cdot \mathbf{n} dS = \int\int_S (P\mathbf{i}) \cdot \mathbf{n} dS + \int\int_S (Q\mathbf{j} \cdot \mathbf{n}) dS + \int\int_S (R\mathbf{k} \cdot \mathbf{n}) dS$. Looking at the third of these integrals, we note that on S_3 the normal direction is perpendicular to the z axis at every point $(x, y, z) \in S_3$ since the surface S_3 contains the vector from $(x, y, h_1(x, y))$ so $(x, y, h_2(x, y))$ passing through (x, y, z). Thus, $(R\mathbf{k} \cdot \mathbf{n}) = 0$ so $\int\int_{S_3} (R\mathbf{k} \cdot \mathbf{n})dS = 0$. Hence, it follows that $\int\int_S (R\mathbf{k} \cdot \mathbf{n})dS = \int\int_{S_2} (R\mathbf{k} \cdot \mathbf{n})dS + \int\int_{S_1} (R\mathbf{k} \cdot \mathbf{n})dS$. Using the usual formula for flux integral through a graph of a function with vector field $< 0, 0, R >$ we have that $= \int\int_{S_2} (R\mathbf{k} \cdot \mathbf{n})dS = \int\int_{D_z} R(x, y, h_2(x, y))dA$ and $= -\int\int_{S_1} (R\mathbf{k} \cdot \mathbf{n})dS = \int\int_{D_z} R(x, y, h_1(x, y))dA$ (since the second integral is oriented downwards). Thus, $\int\int_S (R\mathbf{k} \cdot \mathbf{n})dS = \int\int_{D_z} R(h_2(x, y)) - R(h_1(x, y))dA = \int\int\int_E R_z dV$.

Repeating this argument over the regions D_x, D_y we get that $\int\int_S (P\mathbf{k} \cdot \mathbf{n})dS = \int\int\int_E P_x dV$ and $\int\int_S (Q\mathbf{k}\cdot\mathbf{n})dS = \int\int\int_E Q_y dV$. Thus, we have that $\int\int\int_E \text{div}(\mathbf{F})dV = \int\int_S \mathbf{F} \cdot d\mathbf{S}$. □

As with Green's Theorem, while the most general proof would take us too far afield, we would like to establish the Divergence Theorem for more general regions than simply regions of type one, two and three.

Definition 119

We define a piecewise type one, two and three region inductively as follows. A solid of type one, two and three is a piecewise type one two and three region. If $E = R_1 \cup R_2 \cup ... \cup R_{m-1}$ is a piecewise type one, two and three region where each R_i is a type one two and three region, then if R_m is a type one two and three region so that $E \cap R_m$ is an orientable piecewise smooth graph surface S so that the outward orientation from E on S is opposite the outward orientation from R_m on S, so that the boundary of $E \cup R_m$ is an orientable piecewise smooth graph surface. Further assume that $\partial(E \cup R_m) = \partial(E) \cup \partial(R_m) \setminus I(S)$, where $I(S)$ is the set of points in S which are not elements of the manifold boundary of S. Then $E \cup R_m$ is a *piecewise type one, two and three region.*

Theorem 13.29. *Let E be a piecewise type one, two and three region oriented outwards with boundary surface S. Let $\boldsymbol{F}$ be a C^1 vector field on an open set containing E. Then $\int\int\int_E div(\boldsymbol{F})dV = \int\int_S \boldsymbol{F} \cdot d\boldsymbol{S}$, where S is oriented outwards.*

Proof. We induct on the number of type one two and three regions whose union is E. We have already proven this for a single type one two and three region. Assume that for any $m-1$ type one two and three regions $R_1, ..., R_{m-1}$ it is true that $W = \bigcup_{i=1}^{m-1} R_i$ is a piecewise type one, two and three region so that $\int\int\int_W \text{div}(\mathbf{F})dV = \int\int_T \mathbf{F} \cdot d\mathbf{S}$, where T is the piecewise smooth graph surface which is the boundary of T oriented outwards. Then let $E = W \cup R_m$ for some such W, where R_m is another type one, two and three region so that $W \cap R_m = K$, a piecewise smooth graph surface so that the orientation of K from the outward orientation of W is opposite that of R_m, and the boundary of E is a piecewise smooth orientable surface $S = \partial(R_m) \cup \partial(W) \setminus I(K)$, where $I(K)$ is the set of points of

K which are not on the manifold boundary of K. Then with outward orientation we have $\int\int\int_E \operatorname{div}(\mathbf{F})dV = \int\int\int_W \operatorname{div}(\mathbf{F})dV + \int\int\int_{R_m} \operatorname{div}(\mathbf{F})dV = \int\int_T \mathbf{F}\cdot d\mathbf{S} + \int\int_{\partial(R_m)} \mathbf{F}\cdot d\mathbf{S} =$
$\int\int_{T\setminus I(K)} \mathbf{F}\cdot d\mathbf{S} + \int\int_{\partial(R_m)\setminus I(K)} \mathbf{F}\cdot d\mathbf{S} + \int\int_K \mathbf{F}\cdot d\mathbf{S} - \int\int_K \mathbf{F}\cdot d\mathbf{S} =$
$\int\int_{T\setminus I(K)} \mathbf{F}\cdot d\mathbf{S} + \int\int_{\partial(R_m)\setminus I(K)} \mathbf{F}\cdot d\mathbf{S} = \int\int_S \mathbf{F}\cdot d\mathbf{S}.$

The result follows by induction. □

Essentially, if you can make a solid by sticking a bunch of type one and two and three regions together and gluing them along their boundary surfaces then you can use the Divergence Theorem on that solid. Most solids with connected interiors that you are likely to think of can be assembled that way, which means that the Divergence Theorem works on most solids you are likely to consider.

Here is an example using the Divergence Theorem.

Example 13.22. *Let $\boldsymbol{F}(x, y, z) =< e^z \sin y - 3x, \cos z + 5y, z + \sin(x^3) >$. Find the flux integral $\int\int_S \boldsymbol{F}\cdot d\boldsymbol{S}$, where S is the sphere $x^2 + y^2 + z^2 = 9$ oriented inwards.*

Solution. The sphere bounds a ball E, which is a convex solid, so the Divergence Theorem applies, which states $\iint_{S_1} \mathbf{F}\cdot d\mathbf{S} = \iiint_E \operatorname{div}(\mathbf{F})dV$, if the orientation is outwards. In this case the orientation is inwards and the divergence is $\operatorname{div}(\mathbf{F}) = -3 + 5 + 1 = 3$. Hence, $\iint_{S_1} \mathbf{F}\cdot d\mathbf{S} = -\iiint_E 3dV = -3(\frac{4}{3})\pi(3)^3 = -108\pi.$

□

Exercises:

Exercise 13.1. *The cycloid is the curve traced by* $x(t) = r(t - \sin(t))$ *and* $y(t) = r(1 - \cos(t))$*. One arch of the cycloid is traced out over* $0 \leq t \leq 2\pi$*. Use Green's Theorem to find the area between one arch of the cycloid and the* x*-axis.*

Exercise 13.2. *In addition to defining flux integrals of three dimensional vector fields through a surface we can define a flux integral through a curve oriented in one direction through the curve as follows. Let* C *be a smooth curve* $\boldsymbol{r}(t) = (x(t), y(t))$ *over* $t \in [a, b]$ *in* $\mathbb{R}^2$*. Prove the following:*

(a) If we set $\boldsymbol{n}(t) = \dfrac{(-y'(t), x'(t))}{|r'(t)|}$ *or* $\boldsymbol{n}(t) = \dfrac{(y'(t), -x'(t))}{|r'(t)|}$*. Show* $\boldsymbol{n}(t)$ *is a unit vector which is perpendicular to* $\boldsymbol{r}(t)$*. We refer to the particular choice of* $\boldsymbol{n}(t)$ *as a normal direction through the curve.*

(b) Let $F(x, y) = (P, Q)$ *be a* C^1 *vector field. For a particular choice of* $\boldsymbol{n}(t)$ *as described in (a), if we define the flux of* F *through* C *in the normal direction* $\boldsymbol{n}(t)$ *to* C *to be* $\displaystyle\int_a^b F(\boldsymbol{r}(t)) \cdot \boldsymbol{n}(t)dt$*. Setting* $\boldsymbol{n}(t) = \dfrac{(y'(t), -x'(t))}{|r'(t)|}$*, this is* $\displaystyle\int_C Pdy - Qdx$*. Prove that if* C *is a smooth simple closed curve which oriented counterclockwise which is the boundary of a bounded region* R *then the flux is* $\displaystyle\int_C Pdy - Qdx = \int\int_R div(F)dA$*, and that* $\boldsymbol{n}(t)$ *is pointing outwards from* R *rather than inwards toward* R*.*

Exercise 13.3. *Let* S_1 *be the surface of the cylinder* $x^2 + y^2 \leq 4$ *and* $0 \leq z \leq 4$ *oriented outwards, but without the bottom of the cylinder. Let* $\boldsymbol{F}(x, y, z) =< 3x + \sin(y^2), 4y + ze^{x^2}, 2z >$*. Find the flux integral* $\displaystyle\int\int_{S_1} \boldsymbol{F} \cdot d\boldsymbol{S}$*.*

Exercise 13.4. *Find the area enclosed by the curve* $\boldsymbol{r}(t) =< 2\cos^3(t), 2\sin^3(t) >$ *over* $0 \leq t \leq 2\pi$*.*

Exercise 13.5. *Let* $\alpha(s) =< 3\sin(\frac{s}{5}), 3\cos(\frac{s}{5}), \frac{4s}{5} >$ *be a regular parametrized curve which is parametrized with respect to arc length. Find the curvature and torsion of* $\alpha(s)$*.*

Exercise 13.6. *Let* $\alpha : (-1, 1) \to \mathbb{R}^3$ *be a parametrized curve which is parametrized with respect to arc length and let* $\alpha(0) = s_0$ *and* $\boldsymbol{N} = \boldsymbol{N}(0)$ *be the unit normal to* α *at* s_0*. Show that there if* $\kappa > 0$ *at* s_0 *then there is a* $\delta > 0$ *so that if* $0 < |s| < \delta$ *then* $\alpha(s) \cdot \boldsymbol{N} > 0$*.*

Exercise 13.7. *Let $\boldsymbol{F}(x,y,z) =< y, -x, z-y^2-x^2 >$. Evaluate the flux integral $\displaystyle\int\int_S \boldsymbol{F}\cdot d\boldsymbol{S}$, where S is the surface of the paraboloid $z = x^2+y^2+4$ inside the cylinder $x^2+y^2=1$, oriented upwards.*

Exercise 13.8. *Evaluate the path integral $\displaystyle\int_C \boldsymbol{F}\cdot \boldsymbol{dr}$ where C is the rectangle with vertices $(0,0,2)$, $(2,0,2)$, $(2,2,2)$ and $(0,2,2)$, starting at $(0,0,2)$ and traversed counterclockwise as viewed from above, and*

$$\boldsymbol{F}(x,y,z) =< yz + x^2 + e^{x^2}, xz + 5x + y\sin(y^3), xy + 4x >$$

Exercise 13.9. *Let C be the path consisting of the line segment from $(0,0,2)$ to $(1,5,9)$ followed by the line segment from $(1,5,9)$ to $(8,0,6)$, followed by the line segment from $(8,0,6)$ to $(1,0,3)$. Let $\boldsymbol{F}(x,y,z) =< e^y, xe^y, e^z >$. Find $\displaystyle\int_C \boldsymbol{F}\cdot \boldsymbol{dr}$.*

Exercise 13.10. *Evaluate $\displaystyle\int_C x^2 ds$, where C is the path along the line segment from $(2,0)$ to $(0,4)$.*

Exercise 13.11. *Find the surface area of surface determined by the following parametric equation: $\boldsymbol{r}(u,v) =< u\cos v, u\sin v, v >$, $0 \le u \le 1$, $0 \le v \le \pi$.*

Exercise 13.12. *Prove the following curl form of Green's Theorem: Let C be a positively oriented smooth closed curve which is the boundary of a piecewise type one and type two region E in the xy-plane. Let $\boldsymbol{F} =< P, Q, 0 >$ be a C^1 vector field. Then $\displaystyle\int_C \boldsymbol{F}\cdot \boldsymbol{dr} = \int\int_E (curl(\boldsymbol{F}))\cdot \boldsymbol{k} dA$.*

Exercise 13.13. *Prove the following divergence form of Green's Theorem for flux integrals: Let C be the positively oriented smooth closed curve $\boldsymbol{r} : [a,b] \to \mathbb{R}^2$ bounding the piecewise type one and two region E, and let $F =< P, Q >$ be a C^1 vector field on $\mathbb{R}^2$. Let $\boldsymbol{n} = \dfrac{< y'(t), -x'(t) >}{\sqrt{(x'(t))^2 + (y'(t))^2}}$. Then the flux integral of $\boldsymbol{F}$ through C in direction $\boldsymbol{n}$ is $\displaystyle\int\int_E div(\boldsymbol{F}) dA$.*

Solutions:

Solution to Exercise 13.1. *The cycloid is the curve traced by $x(t) = r(t - \sin(t))$ and $y(t) = r(1 - \cos(t))$. One arch of the cycloid is traced out over $0 \leq t \leq 2\pi$. Use Green's Theorem to find the area between one arch of the cycloid and the x-axis.*

Solution. The area of the region R under one arch is enclosed by the path C consisting of the path C_1 which is $\mathbf{r}(t) = (r(t - \sin(t)), r(1 - \cos(t)))$, $0 \leq t \leq 2\pi$ followed by the path C_2 consisting of $\mathbf{l}(t) = (2\pi r(1-t), 0)$, $0 \leq t \leq 1$. This path is clockwise oriented. If we use the vector field $\mathbf{F}(x, y) =< y, 0 >$ then by Green's Theorem we would have $\int_C \mathbf{F}\cdot d\mathbf{r} = \int\int_R P_y - Q_x dA = \int\int_R 1dA$, which is the area of R. Evaluating $\int_C \mathbf{F}\cdot d\mathbf{r} = \int_{C_1} \mathbf{F}\cdot d\mathbf{r} + \int_{C_2} \mathbf{F}\cdot d\mathbf{r}$, which is $\int_0^{2\pi} < r(1 - \cos(t)), 0 > \cdot < r(1 - \cos(t), r\sin(t) > dt + \int_0^1 < 0, 0 > \cdot < -2\pi r, 0 > dt = \int_0^{2\pi} r^2(1 - 2\cos(t) + \cos^2(t)dt = r^2(2\pi + \pi) = 3\pi r^2$ by Wallis's Formula.

□

Solution to Exercise 13.2. *In addition to defining flux integrals of three dimensional vector fields through a surface we can define a flux integral through a curve oriented in one direction through the curve as follows. Let C be a smooth curve $\boldsymbol{r}(t) = (x(t), y(t))$ over $t \in [a, b]$ in $\mathbb{R}^2$. Prove the following:*

(a) If we set $\boldsymbol{n}(t) = \frac{(-y'(t), x'(t))}{|r'(t)|}$ or $\boldsymbol{n}(t) = \frac{(y'(t), -x'(t))}{|r'(t)|}$. Show $\boldsymbol{n}(t)$ is a unit vector which is perpendicular to $\boldsymbol{r}'(t)$. We refer to the particular choice of $\boldsymbol{n}(t)$ as a normal direction through the curve.

(b) Let $F(x, y) = (P, Q)$ be a C^1 vector field. For a particular choice of $\boldsymbol{n}(t)$ as described in (a), if we define the flux of F through C in the normal direction $\boldsymbol{n}(t)$ to C to be $\int_a^b F(\boldsymbol{r}(t)) \cdot \boldsymbol{n}(t)dt$. Setting $\boldsymbol{n}(t) = \frac{(y'(t), -x'(t))}{|r'(t)|}$, this is $\int_C Pdy - Qdx$. Prove that if C is a smooth simple closed curve which oriented counterclockwise which is the boundary of a bounded region R then the flux is $\int_C Pdy - Qdx = \int\int_R div(F)dA$, and that $\boldsymbol{n}(t)$ is pointing outwards from R rather than inwards toward R.

Solution. (a) We see $\mathbf{n}(t)$ is a unit vector because it is divided by its length. To see $\mathbf{n}(t)$ or orthogonal to $\mathbf{r}'(t)$ we just note that the dot product $\frac{(-y'(t), x'(t))}{|r'(t)|} \cdot (x'(t), y'(t)) = 0$.

(b) Applying Green's Theorem, we see that $\int_C Pdy - Qdx = \int\int_R P_x + Q_y dA = \int\int_R div(\mathbf{F})dA$.

□

Solution to Exercise 13.3. *Let S_1 be the surface of the cylinder $x^2+y^2 \leq 4$ and $0 \leq z \leq 4$ oriented outwards, but without the bottom of the cylinder. Let $\boldsymbol{F}(x,y,z) =< 3x+\sin(y^2), 4y+ze^{x^2}, 2z-2 >=< P, Q, R >$. Find the flux integral $\displaystyle\int\int_{S_1} \boldsymbol{F} \cdot d\boldsymbol{S}$.*

Solution. We will let S_2 by the bottom of the cylinder oriented downwards and $S = S_1 \cup S_2$ and let E be the solid cylinder bounded by S. We note that $\displaystyle\int\int_{S_1} \mathbf{F} \cdot d\mathbf{S} = \int\int_{S} \mathbf{F} \cdot d\mathbf{S} - \int\int_{S_2} \mathbf{F} \cdot d\mathbf{S}$.

Using the Divergence Theorem we know that $\displaystyle\int\int_{S} \mathbf{F} \cdot d\mathbf{S} = \int\int\int_{E} 3+4+2dV = 9(\frac{(4\pi)(4)}{3}) = 48\pi$.

Since S_2 is the graph of $g(x,y) = 0$ over the radius two disk D oriented downwards, we have $\displaystyle\int\int_{S} \mathbf{F} \cdot d\mathbf{S} = \int\int_{D} Pg_x + Qg_y - RdA = \int\int_{D} 0+0+2dA = 8\pi$. Thus, $\displaystyle\int_{S_1} \mathbf{F} \cdot d\mathbf{S} = 48\pi - 8\pi = 40\pi$.

□

Solution to Exercise 13.4. *Find the area enclosed by the curve $\boldsymbol{r}(t) =< 2\cos^3(t), 2\sin^3(t) >$ over $0 \leq t \leq 2\pi$.*

Solution. By symmetry, we note that this is four times the area between the curve and the x-axis over $0 \leq t \leq \frac{\pi}{2}$, and that $x'(t) < 0$ on the interior of this interval. So, the area is $\displaystyle -4\int_0^{\frac{\pi}{2}} 2\sin^3(t)(-6\cos^2(t)(sin(t)dt = 48\int_0^{\frac{\pi}{2}} \sin^4(t)\cos^2(t)dt = 48\left(\frac{(3)(1)(1)}{(6)(4)(2)}\frac{\pi}{2}\right) = \frac{3\pi}{2}$ by Wallis's formula.

□

Solution to Exercise 13.5. *Let $\alpha(s) =< 3\sin(\frac{s}{5}), 3\cos(\frac{s}{5}), \frac{4s}{5} >$ be a regular parametrized curve which is parametrized with respect to arc length. Find the curvature and torsion of $\alpha(s)$.*

Solution. Since $\mathbf{r}$ is parametrized with respect to arc length, the curvature is $\kappa(s) = |\alpha''(s)|$, where $\alpha'(s) =< \frac{3}{5}\cos(\frac{s}{5}), -\frac{3}{5}\sin(\frac{s}{5}), \frac{4}{5} >= T(s)$ and $\alpha''(s) =< -\frac{3}{25}\sin(\frac{s}{5}), -\frac{3}{25}\cos(\frac{s}{5}), 0 >$, which means that $\kappa(s) = \frac{3}{25}$.

To find the torsion, we first note that $\tau(s)\mathbf{N}(s) = \mathbf{B}'(s)$. Since $\mathbf{N}(s) = \frac{\mathbf{T}'(s)}{\kappa(s)}$ we have $\mathbf{N}(s) =< -\sin(\frac{s}{5}), -\cos(\frac{s}{5}), 0 >$. We know that $\mathbf{B}(s) = \mathbf{T}(s) \times \mathbf{N}(s)$, which means that $\mathbf{B}(s) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{3}{5}\cos(\frac{s}{5}) & -\frac{3}{5}\sin(\frac{s}{5}) & \frac{4}{5} \\ -\sin(\frac{s}{5}) & -\cos(\frac{s}{5}) & 0 \end{bmatrix} =< \frac{4}{5}\cos(\frac{s}{5}), -\frac{4}{5}\sin(\frac{s}{5}), -\frac{3}{5} >$, so $\mathbf{B}'(s) =<$

$-\frac{4}{25}\sin(\frac{s}{5}), -\frac{4}{25}\cos(\frac{s}{5}), 0 >$. Since $\tau(s)\mathbf{N}(s) = \mathbf{B}'(s)$, it follows that $\tau(s) = \frac{4}{25}$.

□

Solution to Exercise 13.6. *Let $\alpha : (-1,1) \to \mathbb{R}^3$ be a parametrized curve which is parametrized with respect to arc length and let $\alpha(0) = s_0$ and $\boldsymbol{N} = \boldsymbol{N}(0)$ be the unit normal to α at s_0. Show that there if $\kappa > 0$ at s_0 then there is a $\delta > 0$ so that if $0 < |s| < \delta$ then $\alpha(s) \cdot \boldsymbol{N} > 0$.*

Proof. By Theorem 13.8, if we use the coordinate system where $\alpha(0)$ is the origin and the direction of the x-axis is $\mathbf{T}$, the direction of the y axis is $\mathbf{N}$ and the direction of the z axis is $\mathbf{B}$ where $\mathbf{r}(s) = (x(s), y(s), z(s))$, then there are functions $R_x, R_y, R_z : I \to \mathbb{R}$ so that

(i) $x(s) = s - \frac{\kappa^2}{6}s^3 + R_x$

(ii) $y(s) = \frac{\kappa}{2}s^2 + \frac{\kappa'}{6}s^3 + R_y$

(iii) $z(s) = -\frac{\kappa\tau}{6}s^3 + R_z$

and $\lim\limits_{s\to 0}\frac{R_x}{s^3} = \lim\limits_{s\to 0}\frac{R_y}{s^3} = \lim\limits_{s\to 0}\frac{R_z}{s^3} = 0$

In particular, for some $\delta > 0$, if $|s| < \delta$ then $y(s) = \frac{\kappa}{2}s^2 - \frac{\kappa'}{6}s^3 - R_y > 0$ since both $\frac{\kappa'}{6}s^3$ and R_y have absolute value less than a constant times s^3, which is less than any positive constant times $\frac{\kappa}{2}s^2$ for sufficiently small s. In this coordinate system, $y(s) = \alpha(s) \cdot \mathbf{N}$ so we are finished. □

Solution to Exercise 13.7. *Let $\boldsymbol{F}(x, y, z) =< y, -x, z - y^2 - x^2 >$. Evaluate the flux integral $\int\int_S \boldsymbol{F} \cdot d\boldsymbol{S}$, where S is the surface of the paraboloid $z = x^2 + y^2 + 4$ inside the cylinder $x^2 + y^2 = 1$, oriented upwards.*

Solution. This ia flux integral over a surface that is not the boundary of a solid, so we have no shortcuts. Since the graph of a Cartesian function is the surface, the formula we would use is $\iint_{S_1} \mathbf{F} \cdot d\mathbf{S} = \iint_D -Pg_x - Qg_y + RdA$ (where D is the unit disk), which equals $\iint_D -y(2x) + x(2y) + (x^2 + y^2 + 4 - y^2 - x^2)dA = \iint_D 4 = 4\pi$.

□

Solution to Exercise 13.8. *Evaluate the path integral $\int_C \boldsymbol{F} \cdot \boldsymbol{dr}$ where C is the rectangle with vertices $(0, 0, 2)$, $(2, 0, 2)$, $(2, 2, 2)$ and $(0, 2, 2)$, starting at $(0, 0, 2)$ and traversed counterclockwise as viewed from above, and*

$$\boldsymbol{F}(x, y, z) =< yz + x^2 + e^{x^2}, xz + 5x + y\sin(y^3), xy + 4x >$$

Solution. The path is a closed curve in three dimensions, and the curl of the vector field is simple, so we use Stokes's Theorem. The simplest surface S_1 that this curve bounds appears to be the surface of the graph of $g(x, y) = 2$ over $D = [0, 2] \times [0, 2]$. Since the curve is traversed clockwise as viewed from above, Stokes's Theorem would give us that $\int_C \mathbf{F} \cdot d\mathbf{r} = -\iint_{S_1} \text{curl}(\mathbf{F}) \cdot d\mathbf{S}$, where S_1 is oriented upwards, which is equal to $\iint_D Pg_x + Qg_y - RdA$, where $< P, Q, R >$ represent the coordinates of the curl of $\mathbf{F}$, which is: $\text{curl}(\mathbf{F}) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ z + e^{x^2} & 5x + y\cos(y^2) & z^2 + 4x \end{bmatrix} =< 0, -3, 5 >$. Thus, the flux integral is equal to $\iint_D 0(0) + -3(0) - 5dA = -5(4) = -20.$

□

Solution to Exercise 13.9. *Let C be the path consisting of the line segment from $(0, 0, 2)$ to $(1, 5, 9)$ followed by the line segment from $(1, 5, 9)$ to $(8, 0, 6)$, followed by the line segment from $(8, 0, 6)$ to $(1, 0, 3)$. Let $\boldsymbol{F}(x, y, z) =< e^y, xe^y, e^z >$. Find $\int_C \boldsymbol{F} \cdot \boldsymbol{dr}$.*

Solution. Using the Fundamental Theorem of Line Integrals is faster than integrating directly. We find a potential function f for $\mathbf{F}$ by antidifferentiating the first coordinate with respect to x to get $f = xe^y + g_1(y, z)$, the second with respect to y to get $f = xe^y + g_2(x, z)$ and the third with respect to z to get $f = e^z + g_3(x, y)$. Note that setting $g_1 = e^z$ and $g_2 = e^z$ and $g_3 = xe^y$ gives $f = xe^y + e^z$ in each case, so $\triangledown f = \mathbf{F}$.

Thus, by the Fundamental Theorem of Line Integrals, $\int_C \mathbf{F} \cdot d\mathbf{r} = f(1, 0, 3) - f(0, 0, 2) = 1 + e^3 - e^2$.

□

Solution to Exercise 13.10. *Evaluate $\int_C x^2 ds$, where C is the path along the line segment from $(2, 0)$ to $(0, 4)$.*

Solution. Note that this is a scalar line integral, so we have no shortcuts and must parametrize the line segment C as $\mathbf{r}(t) =< 2 - 2t, 4t >$, $0 \leq t \leq 1$. Then we use the formula $\int_C f ds = \int_a^b f(\mathbf{r}(t))|r'(t)|dt$ to get: $\int_C x^2 ds = \int_0^1 (2 - 2t)^2 \sqrt{4 + 16} dt = 2\sqrt{5} \int_0^1 4t^2 - 8t + 4dt = 2\sqrt{5}(\frac{4}{3}t^3 - 4t^2 + 4t)\Big|_0^1 = \frac{8\sqrt{5}}{3}$.

□

Solution to Exercise 13.11. *Find the surface area of surface determined by the following parametric equation: $\boldsymbol{r}(u, v) =< u\cos v, u\sin v, v >$, $0 \leq u \leq 1$, $0 \leq v \leq \pi$.*

Solution to Exercise 13.12. *Prove the following curl form of Green's Theorem: Let C be a positively oriented smooth closed curve which is the boundary of a piecewise type one and type two region E in the xy-plane. Let $\boldsymbol{F} =< P, Q, 0 >$ be a C^1 vector field. Then $\int_C \boldsymbol{F} \cdot \boldsymbol{dr} = \int\int_E (curl(\boldsymbol{F})) \cdot \boldsymbol{k} dA$.*

Proof. Let $\mathbf{r} =< x(t), y(t), 0 >$ over $a \leq t \leq b$ be a parametrization for C. We know $(\text{curl}(\mathbf{F})) \cdot \mathbf{k} =< R_y - Q_z, P_z - R_x, Q_x - P_y > \cdot \mathbf{k} = Q_x - P_y$. By Green's Theorem, we know that $\int_C \mathbf{F} \cdot \mathbf{r}'(t)dt = \int_a^b < P, Q, 0 > \cdot < x'(t), y'(t), 0 > dt = \int_a^b < P, Q > \cdot < x'(t), y'(t) > dt = \int\int_E Q_x - P_y dA = \int\int_E (\text{curl}(\mathbf{F})) \cdot \mathbf{k} dA$.

□

Solution to Exercise 13.13. *Prove the following divergence form of Green's Theorem for flux integrals: Let C be the positively oriented smooth closed curve $\boldsymbol{r} : [a, b] \to \mathbb{R}^2$ bounding the piecewise type one and two region E, and let $F =< P, Q >$ be a C^1 vector field on $\mathbb{R}^2$. Let $\boldsymbol{n} = \frac{< y'(t), -x'(t) >}{\sqrt{(x'(t))^2 + (y'(t))^2}}$. Then the flux integral of $\boldsymbol{F}$ through C in direction $\boldsymbol{n}$ is $\int\int_E div(\boldsymbol{F}) dA$.*

Proof. By Theorem 13.16, the flux integral of $\mathbf{F}$ through C in direction $\mathbf{n}$ is $\int\int_E P_x + Q_y dA = \int\int_E \text{div}(\mathbf{F}) dA$.

□

Solution. Recall surface area of S_1 is $\iint_D |\mathbf{r}_u \times \mathbf{r}_v| dA$. In this case, $\mathbf{r}_u =< \cos v, \sin(v), 0 >$ and $\mathbf{r}_v =< -u \sin v, u \cos v, 1 >$, so $\mathbf{r}_u \times \mathbf{r}_v = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \cos v & \sin(v) & 0 \\ -u \sin v & u \cos v & 1 \end{bmatrix} =< \sin(v), -\cos(v), u >$ and $|\mathbf{r}_u \times \mathbf{r}_v| = \sqrt{u^2 + 1}$. Thus, $\text{Area}(S_1) = \iint_D |\mathbf{r}_u \times \mathbf{r}_v| dA = \int_0^\pi \int_0^1 \sqrt{u^2 + 1} du dv$. Setting $u = \tan \theta$ we have $du = \sec^2 \theta d\theta$, and thus the integral simplifies to $\pi \int_0^{\frac{\pi}{4}} \sec^3 \theta d\theta =$

$\frac{\pi}{2} (\sec(\theta) \tan(\theta) + \ln |\sec(\theta) + \tan(\theta)|) \Big|_0^{\frac{\pi}{4}} =$
$\frac{\pi}{2} (\sqrt{2} + \ln(\sqrt{2} + 1))$.

□

Chapter 14

Multivariable Supplemental Materials

14.1 Matrices

Definition 120

The (real valued) *matrix* $A = [a_{ij}]_{m\times n}$ is an array of real number entries with m rows and n columns. So, for each $1 \leq i \leq m$ and $1 \leq j \leq n$ we assign a real number a_{ij} to be the entry of A in the ith row and jth column. More formally, we define A to be the function $f_A : \{1, ..., m\} \times \{1, ..., n\} \to \mathbb{R}$ so that $f_A(i, j) = a_{ij}$. If A has m rows and n columns then we say A is an $m \times n$ or "m by n" matrix. Matrix addition is much like vector addition. If $B = [b_{ij}]_{m\times n}$ is another $m \times n$ matrix then we define $\alpha A + \beta B = [\alpha a_{ij} + \beta b_{ij}]_{m\times n}$. If $C = [c_{ij}]_{n\times p}$ is an n by p matrix then the product $AC = [\sum_{k=1}^{n} a_{ik}c_{kj}]_{m\times p}$. When convenient we may denote a row or column of a matrix as being a vector, meaning that the stated vector has the same entries as the corresponding row or column in the same order with respect to the column or row entry order respectively. In particular the jth column vector of a matrix is the vector whose entries are those in the jth column of the matrix, and the jth row vector is the vector whose entries are those in the jth row of the matrix. Using this notation, if A_i is the ith row vector of A and C_j is the jth column vector of C then we can write $AC = [A_i \cdot C_j]_{m\times p}$.

We define $T : \mathbb{R}^n \to \mathbb{R}^m$ to be a *linear transformation* if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$, we have $T(\alpha\mathbf{x}+\beta\mathbf{y}) = \alpha T(\mathbf{x})+\beta T(\mathbf{y})$. Let $\mathbf{e}_j =< 0, 0, 0, ..., 0, 1, 0, 0, ..., 0 >$, the vector whose jth entry is one and whose other entries are zero, called the jth *standard basis vector*.

We also use the notation $\det(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ to denote the determinant of the matrix whose ith row entries are those of the vector $\mathbf{x}_i$.

Definition 121

If $P:\{1,2,3,...,n\}\rightarrow\{1,2,3,...,n\}$ is a one to one and onto function then we refer to it as a *permutation* on the first n natural numbers. If $i<j$ and $P(i)>P(j)$ we refer to $(i,P(i)),(j,P(j))$ as an *inversion* of the permutation and refer to (i,j) as an *inversion pair* for P. We define the permutation $P_{(i,j)}$ that interchanges the ith and jth natural numbers and maps all other natural numbers to themselves to be a *transposition.*

We will refer to a product of entries in an $n\times n$ matrix A that includes exactly one entry factor from each row and column as a permutation product. If $P:\{1,2,3,...,n\}\rightarrow\{1,2,3,...,n\}$ is a permutation then we refer to $\prod\limits_{i=1}^{n}a_{iP(i)}$ as the column permutation product for A corresponding to P and $\prod\limits_{i=1}^{n}a_{P(i)i}$ as the row permutation product corresponding to P. We refer to the (indexed) a_{ij} terms of a permutation product as being *entry factors* of the product. In a given permutation or row or column product for a matrix we say that an *inversion entry pair* is a pair of (indexed) entry factors (a_{ki},a_{rj}) in the product so that $k<r$ and $i>j$.

We first make a few observations about the preceding definition. What we mean by "indexed" in the definition is that the values of the entry factors must be inverted with respect to their index in the matrix (the values of the entries themselves do not determine whether the pair constitutes an inversion). So, the permutation product $a_{12}a_{21}a_{33}$ for a three by three matrix, for example, refers to the product of the terms itself coupled with the associated pairings $(1,2),(2,1),(3,3)$ associated with the permutation. It is the index pairings that determine the inversions corresponding to the product. However, when we refer to operations on permutation products (like adding two permutation products) the output is the non-indexed output corresponding to the operation. So, for instance, the sum of two permutation products is a number (not a number coupled with a collection of indexed pairs).

Next, in a given row permutation product $\prod\limits_{i=1}^{n}a_{P(i)i}$ corresponding to permutation P, a pair $(a_{ki},a_{rj})=(a_{P(k)k},a_{P(r)r})$ is an inversion entry pair if and only if (k,r) is an inversion of P (since being an inversion entry pair corresponds to the ordering of $P(k)$ and $P(r)$ being reversed relative to the order of k and r in the permutation P). Likewise, in the column permutation product $\prod\limits_{i=1}^{n}a_{iP(i)}$ corresponding to permutation P, a pair $(a_{ki},a_{rj})=(a_{kP(k)},a_{rP(r)})$ is an inversion entry pair if and only if (k,r) is an inversion of P.

We next observe that for every row permutation product $\prod\limits_{i=1}^{n}a_{P(i)i}$ for A corresponding to permutation P, there is a unique permutation $Q=P^{-1}$ so that the column permutation product corresponding to Q is $\prod\limits_{i=1}^{n}a_{iQ(i)}$ the same product of the same indexed terms. This is easy enough to see. A row permutation product has exactly one factor entry from each column, so we simply define Q to be the (unique) corresponding permutation that assigns to each column index j the row index of the row i so that $P(i) = j$. In other words, $Q = P^{-1}$.

Associated with the observations above, we notice that the pairs (a_{ks}, a_{rt}) which are row inversions for a given row permutation product $\prod_{i=1}^{n} a_{P(i)i}$ are the same pairs which are inversions for the corresponding column permutation product $\prod_{i=1}^{n} a_{iP^{-1}(i)}$. This is because $k = P(s)$ and $r = P(t)$ and so $s = P^{-1}(r)$ and $t = P^{-1}(t)$. Thus, the order of $P(s)$ and $P(t)$ is reversed relative to the order of s and t if and only if the order of r and k is reversed relative to the order of $P^{-1}(r)$ and $P^{-1}(k)$. It follows that the pairs which are inversions in a given permutation product (and the number of such inversions) is determined by the permutation product terms themselves with their corresponding indices, and is independent of whether the product is considered as a row permutation product or a column permutation product. Thus, in our definition above, what we have called an inversion corresponds to an inversion as a given permutation product is considered as a row permutation product or a column permutation product.

Finally, we observe that the number of inversions in the transposition $P_{(i,j)}$ is $2(j-i)-1$ if $j > i$ (which is an odd number) since each integer between i and j is out of order with respect to the positions of both i and j (giving $i - j - 1$ inversions twice, once for being in reverse order with respect to i and once for j) and also i and j are out of order with respect to each other (but this is just one inversion) in the transposition.

Definition 122

We will define the *sign* $\sigma(P)$ of the permutation P to be 1 if the number of inversions of the permutation is even, and -1 if the number of inversions is odd, or in other words $\sigma(P) = (-1)^m$, where m is the number of inversions of P. We define the *determinant* $\det(A) = |A|$ of an $n \times n$ matrix $A = [a_{ij}]_{n\times n}$ to be $\sum_{P} \sigma(P) \prod_{i=1}^{n} a_{iP(i)}$, where P is understood to range over all permutations of the first n integers.

From this definition, we see that if a matrix has one entry it has one permutation (the identity with zero inversions) and so the determinant is $|a_{11}| = a_{11}$. For a two by two matrix there are two corresponding permutations of column entries by row, either keeping the column entries the same as row entries or interchanging them (corresponding to one inversion) which means that if $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ then $\det(A) = a_{11}a_{22} - a_{21}a_{12}$. While it takes more time to compute, by determining the six possibly permutations of three integers and counting their inversions we get the result that if $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ then $\det(A) = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) - (a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33} + a_{13}a_{22}a_{31})$.

We first observe that whether we permute the row or column entries we arrive at an equivalent definition of determinant.

Theorem 14.1. *Let* $A = [a_{ik}]_{n\times n}$. *Then* $\det(A) = \sum_P \sigma(P) \prod_{i=1}^n a_{P(i)i}$.

Proof. As discussed above, the row permutations products are exactly the column inversion products and their signs are the same. So, $\det(A) = \sum_P \sigma(P) \prod_{i=1}^n a_{P(i)i} = \sum_P \sigma(P) \prod_{i=1}^n a_{iP^{-1}(i)} = \sum_Q \sigma(Q) \prod_{i=1}^n a_{iQ(i)}$. □

Theorem 14.2. *Let* P, Q *be permutations of* $\{1, 2, 3, .., n\}$.

(a) P *is a finite composition of transpositions*

(b) If P *is a composition of a finite sequence of* k *transition permutations then* $\sigma(P) = (-1)^k$.

(c) $\sigma(P \circ Q) = \sigma(P)\sigma(Q)$.

(d) If P *can be written as a composition of an even number of transpositions then* P *cannot be written as a composition of an odd number of transpositions. If* P *can be written as a composition of an odd number of transpositions then* P *cannot be written as a composition of an even number of transpositions.*

Proof. (a) If P is the identity then we apply zero transpositions, which is finite and we are finished. Otherwise, let i_1 be the first integer so that $P(i_1) \neq i_1$. Then $P(i_1) > i_1$. The first transition we apply is $P_{(i_1,P(i_1))}$. This permutation agrees with P on integers $\{1, 2, ..., i_1\}$ since the transition did not change the images of integers preceding i_1. Let i_2 be the first integer greater so that $P(i_2) \neq P_{(i_1,P(i_1))(i_2)}$. Then $P(i_2) > i_2$ and $i_2 > i_1$. We assign the next transition to be performed to be $P_{(i_2,P(i_2))}$. This transition did not move any of the preceding integers on which P agreed with $P_{(i_1,P(i_1))}$ and, in fact, must agree with P on $\{1, 2, 3, ..., i_2\}$. We continue until after a certain number of transpositions (no more than $n - 1$) we have a finite sequence of transpositions whose composition is equal to P.

(b) If we perform a transposition $P_{(i,j)}$ after a permutation P we interchange $P(i)$ and $P(j)$ (we will assume $P(i) < P(j)$). By doing so, for composition $P_{(i,j)} \circ P$, we do not affect inversions for integers paired with images which are not in the interval $[P(i), P(j)]$ because their order relative to the order of their image under $P_{(i,j)} \circ P$ is the same as the relative order under P. However, for any integer w so that $P(i) < P(w) < P(j)$ the relative order of i and w to $P(i)$ and $P(w)$ is reversed. That is, if $i < w$ then $P(i) < P(w)$ would not have created an inversion but now $P_{(i,j)} \circ P(i) = P(j) > P(w)$ so an inversion has been created. Likewise, if $w < i$ then there would have been an inversion previously under P and now there is not under $P_{(i,j)} \circ P$. Either change would change the number of inversions by one. Likewise, the relative order would be reversed with respect with respect to j, meaning that if (w, j) was an inversion under P then it is not under $P_{(i,j)} \circ P$ and if (w, j) was not an inversion under P then it is under $P_{(i,j)} \circ P$. This means that there have been a total of $2(P(j) - P(i))$ inversion changes from integers w so that $P(w) \in (P(i), P(j))$. There is exactly one additional inversion change from the inversions under P and those under $P_{(i,j)} \circ P$, and that is the inversion (or absence of inversion) (i, j) itself. Since that is the only inversion change which is not matched with a corresponding inversion change, the total number of inversion changes is odd (specifically $2(P(j) - P(i)) + 1$), which means that if n_1 is the number of inversions for P and n_2 is the number of inversions for $P_{(i,j)} \circ P$, then the difference is odd, meaning that $\sigma(P_{(i,j)} \circ P) = (-1)^m \sigma(P)$ where m is odd, and therefore $\sigma(P) = -\sigma(P_{(i,j)} \circ P)$.

(c) By (a) we can write P and Q as products of transpositions, each of which multiplies the sign of the permutation by -1 by part (b). Hence, if P is the composition of k transpositions and Q is the product of t transpositions then the $\sigma(P \circ Q) = \sigma(Q \circ P) = (-1)^{k+t} = (-1)^k(-1)^t = \sigma(P)\sigma(Q)$.

(d) By (b) we note that if P is a composition of an even number of transpositions then $\sigma(P) = 1$ and if P is a composition of an odd number of transpositions then $\sigma(P) = -1$. Since the number of inversions in P is a fixed number, and determines $\sigma(P)$ uniquely, is is impossible for P to both be a composition of an even number of transpositions and also an odd number of transpositions.

□

Definition 123

Let $P : \{1, 2, 3, ..., n\} \to \{1, 2, 3, ..., n\}$ be a permutation. If P can be written as a composition of an odd number of transpositions we say that P is an *odd* permutation. If P can be written as a composition of an even number of transpositions then we say that P is an *even* permutation.

Definition 124

The *minor* M_{ij} if the indexed a_{ij} term of a matrix A is the determinant of the matrix obtained by deleting the ith row and jth column of the matrix A. The *cofactor* of a_{ij} is $C_{ij} = (-1)^{i+j} M_{ij}$.

Note that for a two by two matrix, if we move along the first row and multiply each entry by its cofactor and add the resulting products we get the determinant of the matrix. It takes slightly longer to compute it, but this is also true for a three by three matrix. That is, $\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$. This is not a coincidence as we will discuss shortly. We will refer to a sum $\sum_{k=1}^{n} a_{kj}C_{kj}$ for an $n \times n$ matrix as an expansion by cofactors along the jth column, and $\sum_{k=1}^{n} a_{ik}C_{ik}$ as an expansion by cofactors along the ith row.

Theorem 14.3. *Let $A = [a_{ij}]_{n\times n}$ where $n > 1$. Then for any $1 \le i \le n$, $|A| = \sum_{k=1}^{n} a_{ik}C_{ik}$, and for any $1 \le j \le n$ it is true that $|A| = \sum_{k=1}^{n} a_{kj}C_{kj}$.*

Proof. We proceed by induction on n. It is certainly true when $n = 2$ by inspection since we simply multiply the minors (opposite diagonals) to each of the entries and multiply by negative one if the products are along the forward diagonal and -1 along the backwards diagonal and note that each of the four possible expansions does give the determinant. In other words, $a_{11}a_{22} - a_{21}a_{12} = a_{22}a_{11} - a_{12}a_{21} = -a_{21}a_{12} + a_{11}a_{22} = -a_{12}a_{21} + a_{22}a_{11}$.

Next, we assume that the determinant of an $m \times m$ matrix can be obtained by expanding by cofactors along the ith row. We then let $A = [a_{ij}]_{m+1 \times m+1}$ be an $m+1$ by $m+1$ matrix.

To expand along the tth column gives $\displaystyle\sum_{s=1}^{m+1} a_{st}(-1)^{s+t}M_{st}$, and $|A| = \displaystyle\sum_P \sigma(P) \prod_{j=1}^{m+1} a_{jP(j)}$.

For each pair (s,t) the minor $M_{st} = \displaystyle\sum_Q \sigma(Q) \prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$, where Q is a permutation of the first m integers, $b^{(s,t)}_{jQ(j)} = a_{jQ(j)}$ if $j < s$ and $Q(j) < t$, $b^{(s,t)}_{jQ(j)} = a_{j+1Q(j)}$ if $j \geq s$ and $Q(j) < t$, $b^{(s,t)}_{jQ(j)} = a_{j(Q(j)+1)}$ if $j < s$ and $Q(j) \geq t$ and $b^{(s,t)}_{jQ(j)} = a_{(j+1)(Q(j)+1)}$ if $j \geq s$ and $Q(j) \geq t$. We refer to the indexed entry of A equal to the listed $b^{(s,t)}_{jQ(j)}$ as the entry of A corresponding to $b^{(s,t)}_{jQ(j)}$.

The permutation products in the determinant sum $\displaystyle\sum_P \sigma(P) \prod_{j=1}^{m+1} a_{jP(j)}$ which contain a_{st} as an entry factor are exactly the products of the form $b^{(s,t)}_{1Q(1)} b^{(s,t)}_{2Q(2)} ... b^{(s,t)}_{s-1Q(s-1)} a_{st} b^{(s,t)}_{sQ(s)} ... b^{(s,t)}_{mQ(m)}$. Each such product corresponds to a permutation $P : \{1,2,3,...,m+1\} \rightarrow \{1,2,3,...,m+1\}$ where $P(i) = Q(i)$ if $Q(i) < s$, $P(s) = t$, and $P(i) = Q(i) + 1$ if $Q(i) \geq s$, so that $b^{(s,t)}_{1Q(1)} b^{(s,t)}_{2Q(2)} ... b^{(s,t)}_{s-1Q(s-1)} a_{st} b^{(s,t)}_{sQ(s)} ... b^{(s,t)}_{mQ(m)} = a_{1P(1)} a_{2P(2)} ... a_{sP(s)} ... a_{m+1P(m+1)}$.

Next, we wish to determine the difference between the number of inversions in the permutation product $\displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$ and the permutation product $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$. We first notice that if $i < j$ and $Q(i) < Q(j)$ then $P(i) < P(j)$ since $P(j)$ is one of $Q(j)$ or $Q(j) + 1$, and if $P(i) = Q(i) + 1$ then $P(j) = P(j) + 1$. Similarly, if $i > j$ and $Q(i) < Q(j)$ then $P(i) \leq Q(i) + 1 \leq Q(j) \leq P(j)$ and since P is one to one this implies that $P(i) < P(j)$.

Thus, every entry pair $(b^{(s,t)}_{iQ(i)}, b^{(s,t)}_{jQ(j)})$ is an inversion pair for the permutation product $\displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$ if and only if the pair of corresponding entries is an inversion pair for the permutation product $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$.

Since, comparing corresponding entry pairs in the permutation product $\displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$, there are no changes in the inversions in pairs that only include entry factors of $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$ other than a_{st}, the number of inversions of $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$ minus the number inversions of $\displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$ is equal to the number of inversions of the form $(a_{st}, a_{jP(j)})$ for $j \neq s$.

If $i < s$, $a_{iP(i)} = a_{iQ(i)}$ if $1 \le i < t$ and $a_{iP(i)} = a_{iQ(i)+1}$ if $i \ge t$. Thus, if we let k be the number of integers i so that $1 \le i < s$ and $Q(i) \ge t$ then it follows that we get k inversion pairs $(a_{st}, a_{jP(j)})$ where $a_{jP(j)}$ corresponds to $b^{(s,t)}_{iQ(i)}$ for some $i < s$. Likewise, it follows that there are $s - k - 1 = y$ integers i so that $1 \le i < s$ so that $Q(i) < t$ with $a_{jP(j)}$ corresponding to $b^{(s,t)}_{iQ(i)}$, where the $(a_{st}, a_{jP(j)})$ pair is not an inversion pair.

On the other hand, if $i \ge s$ then $b^{(s,t)}_{iQ(i)} = a_{i+1Q(i)}$ if $Q(i) < t$ and $b^{(s,t)}_{iQ(i)} = a_{i+1Q(i)+1}$ if $Q(i) \ge t$. Hence, if $Q(i) < t$ and $a_{jP(j)}$ corresponds to $b^{(s,t)}_{iQ(i)}$ then $(a_{st}, a_{jP(j)})$ is an inversion. If $Q(i) \ge t$ and $i \ge s$ then $i + 1 > s$ and $P(i) = Q(i) + 1 > t$, so the pair $(a_{st}, a_{jP(j)})$ (where $a_{jP(j)}$ corresponds to $b^{(s,t)}_{iQ(i)}$) is not an inversion.

Let r be the number of integers i so that $s \le i \le m$ and $Q(i) < t$. Notice that $r + y = t - 1$ since there are a total of $t - 1$ integers that have images less than t under a permutation since permutations are one to one. Then the total number of inversions entry pairs for the permutation product $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$ containing a_{st} as one element of the pair is $r + k$, which is the difference between the number of inversions in $\displaystyle\prod_{j=1}^{m+1} a_{jP(j)}$ and the number of inversions in $\displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$. Let x denote the number of integers i so that $i \ge s$ and $Q(i) \ge t$. Since there are a total of $m + 1 - t$ integers i so that $Q(i) \ge t$ it follows that $x + k = m + 1 - t$. Likewise, $r + x = m + 1 - s$.

Since $y + k = s - 1$ and $y + r = t - 1$, adding these equations gives $s + t - 2 = 2y + k + r$, so $(s + t) - (k + r) = 2(y + 1)$, which means that the difference between $s + t$ and $k + r$ is even. Hence, $\sigma(P) = (-1)^{k+r}\sigma(Q) = (-1)^{s+t}\sigma(Q)$. This means that we can represent the sum of the signed permutation products containing a_{st} as $(-1)^{s+t} a_{st} \displaystyle\prod_{j=1}^{m} b^{(s,t)}_{jQ(j)}$.

$$\text{Thus, } |A| = \sum_{P} \sigma(P) \prod_{j=1}^{m+1} a_{jP(j)} = \sum_{s=1}^{m+1} a_{st} \sum_{Q} (-1)^{s+t} \sigma(Q) \prod_{j=1}^{m} b^{(s,t)}_{jQ(j)} = \sum_{s=1}^{m+1} (-1)^{s+t} a_{st} M_{st},$$

which is the expansion by cofactors along column t. Similarly, $|A| = \displaystyle\sum_{P} \sigma(P) \prod_{j=1}^{m+1} a_{jP(j)} = \sum_{t=1}^{m+1} a_{st} \sum_{Q} (-1)^{s+t} \sigma(Q) \prod_{j=1}^{m} b^{(s,t)}_{jQ(j)} = \sum_{t=1}^{m+1} (-1)^{s+t} a_{st} M_{st}$, the expansion by cofactors along row s.

□

Definition 125

For a matrix A, the operations of interchanging two rows, adding a multiple of one row to another row and and multiplying a row of A by a non-zero number are called standard *row operation*. The n by n *identity* matrix I is the matrix $[a_{ij}]_{n \times n}$ so that $a_{ij} = 0$ if $i \neq j$ and one if $i = j$. We can also refer to performing row operations on a system of equations listed with first equation then a second below it and so on until an nth equation is listed at the bottom with analogous meanings. Interchanging rows i and j is interchanging ith and jth equations from the top equation, adding a multiple of the ith equation to the jth equation is adding the ith row to the jth row and multiplying the ith row by a non-zero constant refers to multiplying the ith equation by that non-zero constant.

Theorem 14.4. *If a standard row operation is performed on a system of equations then the set of solutions to the system of equations is unchanged.*

Proof. Let the system of equations be as follows:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = c_1\\
&a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = c_2\\
&...\\
&a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = c_m
\end{aligned}$$

A set of entries $x_1, x_2, ..., x_n$ makes the equation true if and only if every equation in the system is true. If so then multiplying both sides of the ith equation by a non-zero number α results in an equation where the equation is still true. Likewise, $\alpha(a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n) = \alpha c_i$ for a non-zero number α then multiplying both sides by $\frac{1}{\alpha}$ the equation original equation is true. Hence, values $x_1, x_2, ..., x_n$ are a solution to the original system if and only if they are solutions to the system where one row is multiplied by a non-zero constant.

Interchanging two equations does not change the system at all, merely the order in which they equations are listed, so this does not alter the set of solutions to the system.

Finally, if $x_1, x_2, ..., x_n$ is a solution to the original system of equations then for any $i \neq j$ in $\{1, 2, 3, ..., m\}$ and $\beta \in \mathbb{R}$ since we know that $a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n = c_i$ and $a_{j1}x_1 + a_{j2}x_2 + ... + a_{jn}x_n = c_j$, it is also true that $\beta(a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n) = \beta c_i$. Thus, $(a_{j1} + \beta a_{i1})x_1 + (a_{j2} + \beta a_{i2})x_2 + ... + (a_{jn} + \beta a_{in})x_2 = c_j + \beta c_i$, so the system of equations where $a_{j1}x_1 + a_{j2}x_2 + ... + a_{jn}x_n = c_j$ is replaced by $(a_{j1} + \beta a_{i1})x_1 + (a_{j2} + \beta a_{i2})x_2 + ... + (a_{jn} + \beta a_{in})x_2 = c_j + \beta c_i$ is true. Likewise, if the system of equations which is the original system except that the equation $a_{j1}x_1 + a_{j2}x_2 + ... + a_{jn}x_n = c_j$ is replaced by $(a_{j1} + \beta a_{i1})x_1 + (a_{j2} + \beta a_{i2})x_2 + ... + (a_{jn} + \beta a_{in})x_2 = c_j + \beta c_i$ is true, since $a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n = c_i$ is true, we know that $-\beta(a_{i1}x_1 + a_{i2}x_2 + ... + a_{in}x_n) = -\beta c_i$ is true. Adding this equation to $(a_{j1} + \beta a_{i1})x_1 + (a_{j2} + \beta a_{i2})x_2 + ... + (a_{jn} + \beta a_{in})x_2 = c_j + \beta c_i$ as before we see that $a_{j1}x_1 + a_{j2}x_2 + ... + a_{jn}x_n = c_j$ is true.

Hence, the system that results from performing a row operation on a given system of equations is always a system with the same solutions as the original system of equations.

□

Theorem 14.5. *Let $A = [a_{ij}]_{n \times n}$ be a matrix.*

(a) If we multiply a row or column of A by a constant k, then the determinant of the resulting matrix is k times the determinant of the original matrix.

(b) If we switch two rows or two columns of A then the determinant of the resulting matrix is the additive inverse of the determinant of the original matrix.

(c) If two rows or columns of A are multiples of each other then the determinant is zero.

(d) If we add a multiple of one row (or column respectively) of A to another row (or column respectively) of A then the determinant of the resulting matrix is the same as the determinant of the original matrix.

Proof. (a) Since each summand in the determinant is a product of terms that includes exactly one term from each row and exactly one term from each column. Hence, multiplying a column or row by k would multiply exactly one factor in each summand by k which would multiply the determinant by k.

(b) We proceed by induction. If $n = 1$ then the result is trivial because there aren't two rows or columns to switch. For $n = 2$, the result is immediate since switching two rows or columns results in determinants $a_{21}a_{12} - a_{22}a_{11}$ and $a_{12}a_{21} - a_{11}a_{22}$ respectively, which is $-|A|$ in each case. Assume the statement is true for a $k \times k$ matrix with $k \geq 2$. Let A be a $k+1 \times k+1$ matrix. If we switch the rth and jth rows of A to achieve a matrix B then since $k+1 \geq 3$ there is another row, say the ith row which was not switched. Expanding by cofactors along the ith row gives us $|B| = \sum_{t=1}^{n} a_{it}(-1)^{(i+t)}M_{it}$, where each M_{it} is the determinant of the corresponding minor in which the rth and jth rows have been switched from the corresponding minor in matrix A. Since each minor was negated by the inductive hypothesis, it follows that the determinant is negated as well. The argument for switching two columns is similar.

(c) If two rows or two columns of A are the same, then switching those rows or columns results in the same matrix which has the same determinant. However, by part (b) the determinant should be negated, meaning $|A| = -|A|$, so $|A| = 0$. By part (a) it follows that if one row is k times another then the determinant of the resulting matrix is $k(0) = 0$ as well.

(d) By replacing a row $\mathbf{r}_i$ of matrix A by $\mathbf{r}_i + k\mathbf{r}_j$ (where $\mathbf{r}_j$ is the jth row of A) and taking the determinant of the resulting matrix B, this changes the entries of each summand in the determinant containing a factor a_{it} by replacing that factor by $a_{it} + ka_{jt}$. This means that the resulting determinant is $\det(A) + \det(A')$, where A' is the matrix obtained by replacing the ith row of A by $k\mathbf{r}_j$, whose determinant is zero by (c). Thus, the determinant of the resulting matrix is the same as that of A. The argument for adding a multiple of a column to another column is similar.

□

Definition 126

The *coefficient matrix* for a system of equations:

$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = c_1$
$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = c_2$
...
$a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = c_n$

is the matrix $A = [a_{ij}]_{n \times n}$.

Theorem 14.6. *Let A be the coefficient matrix for the system of equations:*

$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = c_1$
$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = c_2$
$...$
$a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = c_n$
which is also written $A\boldsymbol{x} = \boldsymbol{c}$, where $\boldsymbol{c} = (c_{1,2}, ..., c_n)$ and $\boldsymbol{x} = (x_1, x_2, ..., x_n)$.

Then $\det(A) \neq 0$ if and only if there is a unique solution to the system. If $\det(A) = 0$ then there are either no solutions to the system of equations or infinitely many solutions to the system.

Furthermore, the determinant of an n by n matrix A is non-zero if and only if there is a sequence of standard row operations that can be performed on A which results in the identity matrix.

Proof. First assume $\det(A) \neq 0$. Since no column of A is a zero vector column (since otherwise the determinant would be zero by 14.3), there is a row, say the j_1th row, which contains a non-zero entry in the first column. Thus, we can add a multiple of row j_1 to the other rows so that the entry of the first column is zero in the resulting matrix in all other rows (and the determinant of the resulting matrix is the same as the original matrix). Then switch the j_1th row with the first row if $j_1 \neq 1$, so that the only entry in the first column which is non-zero is in the first row. This either leaves the determinant the same or negates it. Then divide the first row by its first entry so that the first column consists of a 1 in the first row and a zero in ever other row. The resulting matrix B_1 has determinant of B_1 which is non-zero by Theorem 14.5.

Expanding by cofactors along the first column shows that the determinant of B_1 is the determinant of the matrix obtained by deleting the first row and column of B_1. Thus, since B_1 has non-zero determinant, the second column contains a non-zero entry in some j_2th row where $j_2 > 1$. Add constant multiples of the j_2th row to the other rows to make all entries of the second column zero except for the entry in the j_2th row. Notice that such adding does not affect the first column entries since the first entry in the j_2th row is zero. Then switch the j_2th row with the second row if $j_2 \neq 2$. We then divide the second row by its first non-zero entry to get matrix B_2, where the only non-zero entry of the first column is a one in the first row and the only non-zero entry of the second column is a one in the second row. The matrix B_2 also has non-zero determinant by Theorem 14.5.

Continuing in the manner, we eventually obtain matrix $B_n = I$. Performing the corresponding row operations to the system of equations results in equations $1(x_i) = k_i$ in each row, giving a unique solution to the system of equations.

Next, let $\det(A) \neq 0$. We proceed as before. If there is a non-zero entry in the first column of A then we can create a matrix B_1 as described before, except that $\det(B_1) = 0$ since the row operations performed on A to get B_1 altered the determinant in a manner that multiplied the original determinant of zero by non-zero numbers. If every entry of the first column is zero then there are two possibilities. The first is that there is a solution in the remaining variables $(x_2, x_3, ..., x_n)$ meaning that the following is true.

$a_{12}x_2 + ... + a_{1n}x_n = c_1$
$a_{22}x_2 + ... + a_{2n}x_n = c_2$

...
$a_{n2}x_2 + ... + a_{nn}x_n = c_n$

In that case, if we set x_1 to be any value t then $(t, x_2, x_3, ..., x_n)$ is a solution to the system, so there are infinitely many solutions. If there is no solution to the system listed above (which is the same as the original system) then there are no solutions to the system of equations.

Continuing, if we were able to obtain B_1 then we proceed as before to obtain B_2 unless there is no row below the first row with a non-zero entry in the second column, in which case we let $[b_{ij}^{(1)}]_{n\times n} = B_1$ and let the constant column have entries $b_1^{(1)}, b_2^{(2)}, ..., b_2^{(n)}$ after the row operations performed A to change A to B_1 have been applied to the constant column. Recall from Theorem 14.4, that the solutions to the system $B_1\mathbf{x} = \mathbf{b}^{(1)}$, where $\mathbf{b}^{(1)} = (b_1^{(1)}, b_2^{(2)}, ..., b_2^{(n)})$, are the same as the solutions to the original system of equations. There are again two possibilities. Since there are no non-zero entries below the first row in the second column, either the corresponding system of equations indicated below the first row:
$b_{23}^{(1)}x_3 + ... + b_{2n}^{(1)}x_n = b_2^{(1)}$
$b_{33}^{(1)}x_3 + ... + b_{3n}^{(1)}x_n = b_3^{(1)}$
...
$b_{n3}^{(1)}x_3 + ... + b_{nn}^{(1)}x_n = b_n^{(1)}$
has a solution or it does not. If it does not then there is no solution to the system. If it has a solution $(x_3, x_4, ..., x_n)$ then for every value $t = x_2$ we can back substitute to get $x_1 = b_1^{(1)} - b_{12}^{(1)}t - b_{13}^{(1)}x_3 - ... - b_{1n}^{(1)}x_n$, which gives a solution to the first equation as well, so there are infinitely many solutions.

Now, if we are able to perform a sequence of standard row operations on A and end with $B_n = I$ then by reversing each row operation (switching rows that were switched, adding negative the multiple of rows to other rows for which the original multiple of rows was added to other rows, and multiplying rows by the multiplicative inverse of constants we multiplied them by before) in reverse order we see that a finite sequence of row operations applied to I results in A, which means that $\det(A) \neq 0$ since applying every standard row operation multiplies the determinant by a non-zero number, and the determinant of I is one. This is impossible since $\det(A) = 0$.

We conclude that at some stage in our row reduction process there is some B_j so that we cannot create B_{j+1} as described above because all entries in the $j + 1$st column below the jth row are zero, in which case, as before, we have two cases. Either the corresponding system of equations below the jth row:
$b_{j+1j+2}^{(j)}x_{j+1} + ... + b_{j+1n}^{(j)}x_n = b_{j+1}^{(j)}$
$b_{j+2j+2}^{(j)}x_{j+1} + ... + b_{j+2n}^{(j)}x_n = b_{j+2}^{(j)}$
...
$b_{nj+2}^{(j)}x_{j+1} + ... + b_{nn}^{(j)}x_n = b_n^{(j)}$
has a solution or it does not. If it does not then there is no solution to the system. If it does have a solution then for any value of $t = x_{j+1}$ it follows that by back-substituting to solve for the remaining variables we can find a corresponding solution to the system. In other words, by setting:
$x_j = b_j^{(j)} - b_{jj+1}^{(j)}t - b_{jj+2}^{(j)}x_{j+2} - ... - b_{jn}^{(j)}x_n$
$x_{j-1} = b_{j-1}^{(j)} - b_{j-1j}^{(j)}x_j - b_{jj+1}^{(j)}t - ... - b_{j-1n}^{(j)}x_n$
...

$x_1 = b_1^{(j)} - b_{12}^{(j)}x_2 - b_{13}^{(1)}x_3 - ... - b_{jj+1}^{(j)}t - ... - b_{1n}^{(j)}x_n$
we have a solution to the system. Hence, there are infinitely many solutions to the system of equations.

□

Theorem 14.7. *Cramer's Rule. Let*
$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = c_1$
$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = c_2$
$\dots$
$a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = c_n$
be a system of equations with coefficient matrix A having non-zero determinant. Let A_{x_i} be the matrix A with the ith row replaced by the constant column vector $(c_1, c_2, ..., c_n)$. Then $x_i = \dfrac{|A_i|}{|A|}$.

Proof. We know the system of equations has a solution by Theorem 14.6. Then it is true that

$$\begin{vmatrix} a_{11} & a_{12} & ... & a_{1(i-1)} & a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n & a_{1(i+1)} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2(i-1)} & a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n & a_{2(i+1)} & ... & a_{2n} \\ .... \\ a_{n1} & a_{n2} & ... & a_{n(i-1)} & a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n & a_{n(i+1)} & ... & a_{nn} \end{vmatrix}$$

$= \begin{vmatrix} a_{11} & a_{12} & ... & c_1 & ... & a_{1n} \\ a_{21} & a_{22} & ... & c_2 & ... & a_{2n} \\ .... \\ a_{n1} & a_{n2} & ... & c_n & ... & a_{nn} \end{vmatrix} = |A_i|$. However, by Theorem 14.5 it also follows that this is the same as $\begin{vmatrix} a_{11} & a_{12} & ... & x_ia_{1i} & ... & a_{1n} \\ a_{21} & a_{22} & ... & x_ia_{2i} & ... & a_{2n} \\ .... \\ a_{n1} & a_{n2} & ... & x_ia_{ni} & ... & a_{nn} \end{vmatrix} = x_i|A|$. Thus, $x_i = \dfrac{|A_i|}{|A|}$.

□

Theorem 14.8. *Let $A = [a_{ij}]_{m\times n}$ be a matrix and $\mathbf{e}_j$ be the jth standard basis vector A_j of $\mathbb{R}^n$. Then $A\mathbf{e}_j = A_j$, the column matrix whose entries are those of the jth column vector. If B is a matrix so that $B\mathbf{e}_j = A_j$ for all $1 \le j \le n$ then $A = B$.*

Proof. By definition of matrix multiplication, $A\mathbf{e}_j = [\sum_{k=1}^{n} a_{ik}\mathbf{e}_{j_k}]_{1\times m}$ where $\mathbf{e}_{j_k}$ is the kth entry of $\mathbf{e}_j$. Since all entries except the jth entry are zero and the jth entry is one, this is equal to $[a_{ij}]_{1\times m} = A_j$. If we replace A by B then since every column of B is the corresponding column of A we know $A = B$. □

Theorem 14.9. *Let $A = [a_{ij}]_{m\times n}$ be a matrix. Then the function $T(\boldsymbol{x}) = A\boldsymbol{x}$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$.*

Proof. First, the fact that each point in $\mathbb{R}^n$ can be mapped under this definition, and that the range would be in $\mathbb{R}^m$ follows directly from the definition of matrix multiplication. To show linearity, note that $A(\alpha\mathbf{x}+\beta\mathbf{y}) = [\sum\limits_{k=1}^{n} a_{ik}(\alpha x_k+\beta y_k)]_{1\times m} = \alpha[\sum\limits_{k=1}^{n} a_{ik}x_k]_{1\times m} + \beta[\sum\limits_{k=1}^{n} a_{ik}y_k]_{1\times m}$. □

Theorem 14.10. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then there is exactly one matrix $A = [a_{ij}]_{m\times n}$ so that $T(\boldsymbol{x}) = A\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{R}^n$.*

Proof. Define $A_j = T(\mathbf{e}_j)$ where A_j is the jth column vector of A. Then by Theorem 14.8, $T(\mathbf{e}_j) = A\mathbf{e}_j$ for each $1 \le j \le n$. Since $A\mathbf{x}$ is also linear by Theorem 14.9, for any vector $\mathbf{x} =< x_1, x_2, x_3, ..., x_n >\in \mathbb{R}^n$ we have that $T(\mathbf{x}) = x_1T(\mathbf{e}_1)+x_2T(\mathbf{e}_2)+...+x_nT(\mathbf{e}_n) = A\mathbf{x}$. Since we must have $A_j = T(\mathbf{e}_j)$ for $A\mathbf{e}_j$ to be equal to $T(\mathbf{e}_j)$, the choice of matrix A is unique. □

Theorem 14.11. *Let $A = [a_{ij}]_{m\times n}$, $B = [b_{ij}]_{n\times r}$, $C = [c_{ij}]_{r\times t}$. Then $AB(C) = (AB)C$.*

Proof. There are unique linear transformations T_A, T_B, T_C so that $T_C(\mathbf{x}) = C\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^t$, $T_B(\mathbf{x}) = B\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^r$, and $T_A(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$. We also know that $T_A(T_B \circ T_C(\mathbf{x})) = (T_A \circ T_B)(T_C(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^t$. This means that the matrices corresponding to these function compositions are the same. Since $A(BC\mathbf{x}) = T_A(T_B{\circ}T_C)(\mathbf{x})$ and $(AB)C\mathbf{x} = T_A \circ T_B(T_C(\mathbf{x}))$, it follows that $AB(C) = (AB)C$.

Alternately, by definition $AB = [\sum\limits_{k=1}^{n} a_{ik}b_{kj}]_{m\times r}$ and $(AB)C = [\sum\limits_{s=1}^{r}(\sum\limits_{k=1}^{n} a_{ik}b_{ks})c_{sj}]_{m\times t}$. On the other hand, $BC = [\sum\limits_{s=1}^{r} b_{is}c_{sj}]_{n\times t}$, which means that $A(BC) = [\sum\limits_{k=1}^{n} a_{ik}\sum\limits_{s=1}^{r} b_{ks}c_{sj}]_{m\times t}$. Since these are equal, matrix multiplication is associative as desired. □

Definition 127

We say that an $n \times n$ matrix A has an *inverse* A^{-1} if $AA^{-1} = A^{-1}A = I$. We say that A is *invertible* if there is an inverse matrix for A.

Theorem 14.12. *Let A be an $n \times n$ matrix so that for some matrix B it is true that $AB = I$. Then $BA = I$ and $B = A^{-1}$.*

Proof. Since $AB = I$ by Theorem 14.11, we know that for any vector $\mathbf{x} \in \mathbb{R}^n$ it is true that $AB\mathbf{x} = I\mathbf{x} = \mathbf{x}$. This means that the linear transformations $A\mathbf{x}$ and $B\mathbf{x}$ are inverses of one another. However, for any functions, the composition of a function with its inverse in either order is the identity. Hence, $BA\mathbf{x} = \mathbf{x}$, which means that $BA\mathbf{x} = I\mathbf{x}$ which implies that $BA = I$ by Theorem 14.10 □

Definition 128

Let $E_{(i,j)}$ be the notation for the matrix obtained by switching the ith row and jth row of the identity matrix. This is the elementary matrix associated with the row operation of switching rows i and j.

Let $E_i^{(\alpha)}$ be the notation of the matrix obtained by replacing the 1 entry in the ith row of the identity matrix by the non-zero number α. This is the elementary matrix associated with the row operation of multiplying the ith row by α.

Let $E_{(i,j)}^{(\alpha)}$ be the matrix obtained by adding α times the ith row of the identity matrix to the jth row of the identity matrix. This is the elementary matrix associated with the row operation of adding α times the ith row of A to the jth row of A.

Matrices of these three types are called *elementary* matrices.

Theorem 14.13. *(a) Let $A = [a_{ij}]_{n\times n}$ be an n by n matrix and let $i, j \in \{1, 2, 3, ..., n\}$. Then $E_{(i,j)}A$ is the matrix obtained by switching the ith and jth rows of A, $E_{(i,j)}^{(\alpha)}A$ is the matrix obtained by adding α times the ith row of A to the jth row of A, and $E_i^{(\alpha)}A$ is the matrix obtained by multiplying the ith row of A by α.*

(b) The matrix obtained by performing any finite sequence of row operations on A is the product of the corresponding elementary matrices outlined in part (a) times A.

(c) The determinant of the product of elementary matrices is the product of the determinants of these matrices, which is always non-zero.

(d) Every product of elementary matrices is invertible.

(e) If the determinant of a A is non-zero then the A is the product of elementary matrices.

(f) The matrix A is invertible if and only if $\det(A) \neq 0$.

(g) The equation the linear transformation $A\boldsymbol{x} = \boldsymbol{0}$ has a unique solution of $\boldsymbol{x} = \boldsymbol{0}$ if and only if $A\boldsymbol{x}$ is both one to one and onto, which is true if and only if $\det(A) \neq 0$.

(h) If A and B are n by n matrices then $\det(AB) = \det(A)\det(B)$.

(i) A matrix is invertible if and only if it is a product of elementary matrices.

Proof. (a) First, observe that $E_{(i,j)}A$ has a one in the jth entry of the ith row and the other entries are zeroes. Thus, the kth entry of the ith row of the product matrix is $\mathbf{e}_j \cdot A_k$, where A_k is the kth column of A, which is equal to a_{jk}. Similarly, the kth entry of the jth row of the product is $\mathbf{e}_i \cdot A_k = a_{ik}$. For any other $1 \leq m \leq n$ the only non-zero entry of the mth row of $E_{(i,j)}$ is 1 in the mth entry, so the kth entry of the mth row is a_{mk}.

Next, note that as above, all entries of rows of $E_i^{(\alpha)}$ other than the ith row are the same as those of the identity matrix, so all rows of $E_i^{(\alpha)}A$ are the same as those of A except for the ith row. In the ith row all entries are zero except for an α in the ith position, so $\alpha\mathbf{e}_i \cdot A_k = \alpha a_{ik}$ is the kth entry of the ith row of the product for each $1 \leq k \leq n$.

Finally, as above, all entries of rows of $E_{(i,j)}^{(\alpha)}$ other than the ith row are the same as those of the identity matrix, so all rows of $E_i^{(\alpha)}A$ are the same as those of A apart from the ith row. In the ith row all rows as zero except for a 1 in the ith position and an α in the jth position. So, by definition of matrix multiplication, the kth entry of the ith row of the product is $(\mathbf{e}_i + \alpha\mathbf{e}_j) \cdot A_k = a_{ik} + \alpha a_{jk}$.

(b) It follows from (a) that performing a sequence of row operations results in the same matrix as multiplying by the associated row operations on the left sequentially, which is the same as multiplying by the matrix which is the product of such elementary matrices by Theorem 14.11.

(c) We know from Theorem 14.5 that switching rows of a matrix negates the determinant, multiplying a row by α multiples the determinant by α and adding a multiple of a row to another row does not change the determinant (it multiplies the determinant by one). The determinant of I is one, so $\det(E_{(i,j)}I) = \det(E_{(i,j)}) = -1$, $\det(E_{(i,j)}^{(\alpha)}I) = \det(E_{(i,j)}^{(\alpha)}) = 1$, and $\det(E_i^{(\alpha)}I) = \det(E_i^{(\alpha)}) = \alpha \neq 0$ for elementary matrices $E_{(i,j)}, E_i^{(\alpha)}, E_{(i,j)}^{(\alpha)}$. Performing these operations in sequence on I causes the multiplication by $-1, 1$ or α respectively in sequence to the resulting determinant, which means that the determinant of the product of elementary matrices is the product of their determinants. Since each determinant of each elementary matrix is non-zero, the determinant of the product is also non-zero.

(d) Each elementary matrix is invertible. In particular $E_{(i,j)}^{-1} = E_{(i,j)}$, the inverse of $E_i^{(\alpha)}$ is $E_i^{(\frac{1}{\alpha})}$ and the inverse of $E_{(i,j)}^{(\alpha)}$ is $E_{(i,j)}^{(-\alpha)}$. Hence, for any product $B = E_1E_2E_3...E_k$ of elementary matrices, the product $E_k^{-1}E_{k-1}^{-1}...E_2^{-1}E_1^{-1} = B^{-1}$.

(e) Let $\det(A) \neq 0$. Then we know by Theorem 14.6 that we can perform a sequence of row operations on A to change A to I. By part (b) we know that performing this sequence of row operations is the same as multiplying on the left the sequence of elementary matrices associated with these row operations. So, if changing A to I is done by a sequence of operations whose associated elementary matrices are $E_1, E_2, ..., E_k$, then $E_kE_{k-1}...E_2E_1A = I$. Thus, $A^{-1} = E_kE_{k-1}...E_2E_1$ by Theorem 14.12, which means that $A = E_1^{-1}E_2^{-1}...E_n^{-1}$.

(f) If $\det(A) \neq 0$ then we know that A can be converted to I by performing a finite sequence of row operations by Theorem 14.6. Hence, if these row operations are associated with multiplying by elementary matrices $E_1, E_2, ..., E_m$, then $E_mE_{m-1}...E_1A = I$, so $E_mE_{m-1}...E_1 = A^{-1}$. If A is invertible then if $A\mathbf{x} = \mathbf{0}$ it must follow that $A^{-1}A\mathbf{x} = A^{-1}\mathbf{0} = 0$, so $\mathbf{x} = \mathbf{0}$. Since the solution to this is unique, by Theorem 14.6, again, we know that $\det(A) \neq 0$.

(g) If $A\mathbf{x} = \mathbf{0}$ has only the solution $\mathbf{x} = \mathbf{0}$ then if $A\mathbf{x} = A\mathbf{y}$ for some $\mathbf{x}, \mathbf{y}$ we know that $A(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ which means that $\mathbf{x} - \mathbf{y} = \mathbf{0}$ so $\mathbf{x} = \mathbf{y}$, so the linear transformation $T(\mathbf{x}) = A\mathbf{x}$ is one to one. Since the solution is unique, we also know that the determinant of A is non-zero which implies that for any vector $\mathbf{c} \in \mathbb{R}^n$ there is a unique $\mathbf{x}$ so that $A\mathbf{x} = \mathbf{c}$ by Theorem 14.6. This means that f is both one to one and onto. This is also equivalent to $\det(A) \neq 0$ by Theorem 14.6 and part (b).

(h) In (c), (e) we have established that the determinant of a matrix is non-zero if and only if it is the product of elementary matrices. Hence, if $\det(A) \neq 0$ and $\det(B) \neq 0$ then we know that $A = A_1A_2...A_k$ for some elementary matrices $A_1, A_2, ..., A_k$ and $B = B_1B_2...B_t$ for elementary matrices $B_1B_2...B_t$. By (c), $\det(AB) = \det(A_1A_2...A_kB_1B_2...B_t) = (\det(A_1)\det(A_2)...\det(A_k))(\det(B_1\det(B_2)...\det(B_k)) = \det(A)\det(B)$. If $\det(B) = 0$ then we know that $B\mathbf{x} = \mathbf{0}$ has infinitely many solutions by Theorem 14.6 (since we know that $\mathbf{x} = \mathbf{0}$ is a solution), so it has a non-zero solution $\mathbf{p}$. Hence, $AB\mathbf{p} = A\mathbf{0} = \mathbf{0}$ which means that $AB\mathbf{x} = 0$ has a non-zero solution, so $\det(AB) = 0$. If $\det(B) \neq 0$ and $\det(A) = 0$ then there is some $\mathbf{q} \neq \mathbf{0}$ so that $A\mathbf{q} = 0$, and since B is invertible it follows that $ABB^{-1}\mathbf{q} = A\mathbf{q} = 0$. Thus, $\det(AB) = 0$.

(i) This follows from (d), (e), (c), and (f).

□

14.2 Partitions of Unity

Partitions of unity were not needed for any of the results in this text, but we include a development of them here because they are useful in Advanced Calculus for extending local properties to global ones. The format of this development largely parallels that of Spivak [1].

Theorem 14.14. *Let $f(x) = e^{-\frac{1}{x^2}}$ if $x \neq 0$ and let $f(0) = 0$. Prove that f has derivatives of all orders, but is not analytic on any open interval containing 0. Furthermore, the functions $g(x) = e^{-\frac{1}{x^2}}$ if $x > 0$ and $g(x) = 0$ if $x \leq 0$ and $h(x) = e^{-\frac{1}{x^2}}$ if $x < 0$ and $h(x) = 0$ if $x \geq 0$ are C^∞, and $f^{(n)}(0) = g^{(n)}(0) = h^{(n)}(0) = 0$ for every natural number n.*

Proof. Each derivative of $f^{(n)}(x)$ is a rational function multiplied by $e^{\frac{-1}{x^2}}$ everywhere except at $x = 0$. To see this, note that this is true for $n = 0$ and if the kth derivative is $\frac{p(x)}{q(x)}e^{\frac{-1}{x^2}}$ then $f^{(k+1)}(x) = (\frac{p(x)}{q(x)}\frac{2}{x^3} + \frac{q(x)p'(x) - p(x)q'(x)}{(q(x))^2})e^{-\frac{1}{x^2}}$, so the statement follows inductively.

At $x = 0$, let $f^{(n)}(x) = \frac{p(x)}{q(x)}e^{-\frac{1}{x^2}}$ for $x \neq 0$. Then we have $f^{(n)}(0) = \lim_{x\to 0} \frac{\frac{p(x)}{q(x)}e^{-\frac{1}{x^2}} - 0}{x - 0}$ $= \frac{p(x)}{xq(x)}e^{-\frac{1}{x^2}}$, which is a rational function times $e^{-\frac{1}{x^2}}$. For any positive integer m is true that $\lim_{x\to 0^+} \frac{1}{x^m}e^{-\frac{1}{x^2}} = \lim_{u\to\infty} u^m e^{-u^2}$ by Theorem 4.14. By L'Hospital's rule we know that $\lim_{u\to\infty} \frac{u^m}{e^u} = 0$ since differentiating the numerator m times leaves a derivative of $m!$ and differentiating the denominator still leaves e^u and $\lim_{u\to\infty} \frac{m!}{e^u} = 0$. Since $u^m e^{-u^2} < \frac{u^m}{e^u}$ for all $u > 1$ it follows that $\lim_{u\to\infty} u^m e^{-u^2} = 0$ by the Squeeze Theorem for extended real numbers. Similarly, it follows that $\lim_{x\to 0^-} \frac{1}{x^m}e^{-\frac{1}{x^2}} = 0$.

If we choose m larger than the degree of $xq(x)$ then $\lim_{x\to 0} \frac{x^m}{xq(x)} = 0$ since if k is the power of the lowest power summand Cx^k in $xp(x)$ then dividing the numerator and denominator by x^k gives us a numerator x^{m-k} which approaches zero, and a denominator which approaches C as x approaches zero. Since $0 < \frac{1}{xq(x)}e^{\frac{-1}{x^2}} < \frac{1}{x^m}e^{-\frac{1}{x^2}}$ for x sufficiently close to zero, it follows from the Squeeze Theorem that $\lim_{x\to 0} \frac{1}{xq(x)}e^{-\frac{1}{x^2}} = 0$. We also know that $\lim_{x\to 0} p(x) = p(0)$. Thus, by the product rule for limits $f^{(n)}(0) = \lim_{x\to 0} \frac{p(x)}{xq(x)}e^{-\frac{1}{x^2}} = p(0)(0) = 0$.

Thus, all derivatives of all orders for $f(x)$ are zero at $x = 0$, so the Maclaurin series for $f(x)$ is valid only at a single point.

To see that the function g is C^∞ we first note that derivatives of all orders are zero at points $x < 0$ and derivatives of all orders exist at points $x > 0$ since $g(x) = f(x)$ on $(0, \infty)$. At 0 we observe that for each natural number n it is true that $\lim_{x\to 0^-} \frac{g^{(n)}(x) - g^{(n)}(0)}{x - 0} = \frac{0 - 0}{x} = 0$ and $\lim_{x\to 0^+} \frac{g^{(n)}(x) - g^{(n)}(0)}{x - 0} = f^{(n+1)}(0) = 0$. Since $h(x) = g(-x)$, it also follows that h is C^∞ and $h^{(n)}(0) = 0$ for every natural number n by the chain rule. □

Definition 129

The *support* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is $spt(f) = \overline{f^{-1}(\mathbb{R} \setminus \{0\})}$.

We note that since the support of a function is closed, it is compact if and only if it is bounded.

We next demonstrate that there is a C^∞ function that is positive on a given open interval and zero outside of that interval.

Theorem 14.15. *For any interval (a, b), there is a function $w_{(a,b)} : \mathbb{R} \to \mathbb{R}$ which is C^∞ so that $w_{(a,b)}(x) > 0$ for all $x \in (a, b)$ and $w_{(a,b)}(x) = 0$ otherwise.*

Proof. We set $w_{(a,b)}(x) = e^{\frac{-1}{(x-a)^2}} e^{\frac{-1}{(x-b)}^2}$ if $x \in (a, b)$ and $w_{(a,b)}(x) = 0$ otherwise. Since this is just the product $g(x-a)h(x-b)$ from Theorem 14.14, we know that w is C^∞ (by the product rule). □

We next observe that we can find a C^∞ function for a given open interval (a, b) that is zero until the left end point of the interval, and one after the right end point of of the interval and positive in between.

Theorem 14.16. *Let (a, b) be an open interval and let $w_{(a,b)} : \mathbb{R} \to \mathbb{R}$ be a C^∞ function so that $w_{(a,b)}(x) > 0$ if $x \in (a, b)$ and $w_{(a,b)}(x) = 0$ otherwise. Then there is a C^∞ function $h_{(a,b)}$ so that $h_{(a,b)}(x) = 0$ if $x \le a$ and $h_{(a,b)}(x) = 1$ if $x \ge b$ and $h_{(a,b)}(x) > 0$ if $a < x < b$.*

Proof. We define $h_{(a,b)}(x) = \dfrac{\int_a^x w_{(a,b)}(t)dt}{\int_a^b w_{(a,b)}(t)dt}$. By the Fundamental Theorem of Calculus we see that $h'_{(a,b)}(x) = \dfrac{w_{(a,b)}(x)}{\int_a^b w_{(a,b)}(t)dt}$ is C^∞, so $h_{(a,b)}$ is C^∞. If $x \le a$ then since $w_{(a,b)}(t) = 0$ for all $t \in [x, a]$, we know that $\displaystyle\int_a^x w_{(a,b)}(t)dt = 0$, so $h_{(a,b)}(x) = 0$ as well. If $x \in (a, b)$ then since $w_{(a,b)}$ is continuous and positive we know that $h_{(a,b)}(x) > 0$ by Exercise 6.5. If $x \ge b$ then $h_{(a,b)}(x) = \dfrac{\int_a^x w_{(a,b)}(t)dt}{\int_a^b w_{(a,b)}(t)dt} = \dfrac{\int_a^b w_{(a,b)}(t)dt}{\int_a^b w_{(a,b)}(t)dt} + \dfrac{\int_b^x w_{(a,b)}(t)dt}{\int_a^b w_{(a,b)}(t)}dt = 1 + 0 = 0$ since $w_{(a,b)}(t) = 0$ for all $t \in [b, x]$. □

The next step is to move into $\mathbb{R}^n$ by showing that we can find a function which is C^∞ on a cube centered at a point which is one near the point, positive a little further from the point and then zero outside of a small cube about the point. Since the notion of an ϵ cube is useful in more than one context, we will just define a notation for this idea first.

Theorem 14.17. *Let $\boldsymbol{p} = (p_1, p_2, p_3, ..., p_n) \in \mathbb{R}^n$, and let $\epsilon > 0$. Then there is a C^∞ function $g_{(\boldsymbol{p},\epsilon)} : \mathbb{R}^n \to \mathbb{R}$ so that $g_{(\boldsymbol{p},\epsilon)}(\boldsymbol{x}) > 0$ if $\boldsymbol{x} \in R_\epsilon(\boldsymbol{p})^\circ$ and $g_{(\boldsymbol{p},\epsilon)}(\boldsymbol{x}) = 0$ otherwise, and $spt(g_{(\boldsymbol{p},\epsilon)}) = R_\epsilon(\boldsymbol{p})$.*

Proof. Let $w_{(a,b)}$ be as described in Theorem 14.15. For any $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$, we define $g_{(\mathbf{p},\epsilon)}(\mathbf{x}) = \displaystyle\prod_{i=1}^{n} w_{(p_i-\epsilon,p_i+\epsilon)}(x_i)$. Then $g_{(\mathbf{p},\epsilon)}$ is C^∞ since each $w_{(p_i-\epsilon,p_i+\epsilon)}$ is C^∞. If $\mathbf{x} \displaystyle\int R_\epsilon(\mathbf{p})^\circ$ then $p_i-\epsilon < x_i < p_i+\epsilon$ for each i, which means that $g_{(\mathbf{p},\epsilon)}(\mathbf{x}) > 0$. If $\mathbf{x} \notin R_\epsilon(\mathbf{p})^\circ$ then for some i it must be true that $x_i \notin (p_i-\epsilon, p_i+\epsilon)$ which means that $w_{(p_i-\epsilon,p_i+\epsilon)}(x_i) = 0$, so $g_{(\mathbf{p},\epsilon)}(\mathbf{x}) = 0$. Since $\overline{R_\epsilon(\mathbf{p})^\circ} = R_\epsilon(\mathbf{p})$ this means that $spt(g_{(\mathbf{p},\epsilon)}) = R_\epsilon(\mathbf{p})$. □

In the field of topology, there is a theorem called Urysohn's Lemma which shows that for a normal topological space and any two disjoint closed sets in that space, there is a continuous function from the space to the interval $[0, 1]$ which takes one closed set to zero and the other to one. The following theorem can be thought of as an analog to this lemma in $\mathbb{N}$ for functions which are C^∞ where one of the two closed sets is compact.

Theorem 14.18. *Let K be a compact subset of $\mathbb{R}^n$ and let U be an open set containing K. Then there is a C^∞ function $f_{(K,U)} : \mathbb{R}^n \to \mathbb{R}$ and an open V so that $K \subset V \subset \overline{V} \subset U$ and $f_{(K,U)}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in K$, $0 < f_{(K,U)}(\boldsymbol{x}) \leq 1$ if $\boldsymbol{x} \in V$ and $f_{(K,U)}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \notin V$, so $spt(f_{(K,U)}) = \overline{V} \subset U$.*

Proof. Let $g_{(\mathbf{x},\epsilon)}$ and $h_{(a,b)}$ be as defined in theorems 14.17 and 14.16.

By the Lebesgue Number Lemma we can find an $\epsilon > 0$ so that if the diameter of a set S is less than or equal to 2ϵ and S intersects K then $S \subset U$.

Let $C = \{R_\epsilon(\mathbf{x})^\circ\}_{\mathbf{x}\in K}$. Since K is compact there is a finite $F \subseteq C$ which covers K, where $F = \{R_\epsilon(\mathbf{x}_i)^\circ\}_{1\leq i\leq m}$. Define $f(\mathbf{x}) = \displaystyle\sum_{i=1}^{m} g_{(\mathbf{x}_i,2\epsilon)}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. Set $V = \displaystyle\bigcup_{i=1}^{m} R_{2\epsilon}(\mathbf{x}_i)^\circ$. Then $K \subset \displaystyle\bigcup_{i=1}^{m} R_\epsilon(\mathbf{x}_i)^\circ \subset V \subset \overline{V} \subset U$.

If $\mathbf{x} \in V$ then for some i it follows that $\mathbf{x} \in R_{2\epsilon}(\mathbf{x}_i)^\circ$ which means that $f(\mathbf{x}) \geq g_{(\mathbf{x}_i,2\epsilon)}(\mathbf{x}) > 0$ since $\mathbf{x} \in R_{2\epsilon}(\mathbf{x}_i)^\circ$. If $\mathbf{x} \notin V$ then $x \notin R_{2\epsilon}(\mathbf{x}_i)^\circ$ for any natural number $i \leq m$, which means that $g_{(\mathbf{x}_i,2\epsilon)}(\mathbf{x}) = 0$ for each i, so $f(\mathbf{x}) = 0$. By the Extreme Value Theorem, since K is compact and f is positive on K, we know that f takes on a minimum value r on K.

We define $f_{(K,U)}(\mathbf{x}) = h_{(0,r)}(f(\mathbf{x}))$. Then for all $\mathbf{x} \in K$ we know that $f(\mathbf{x}) \geq r$, so $f_{(K,U)}(\mathbf{x}) = h_{(0,r)}(f(\mathbf{x})) = 1$, and for all $\mathbf{x} \notin V$, we know that $f(\mathbf{x}) = 0$, so $f_{(K,U)}(\mathbf{x}) = h_{(0,r)}(f(\mathbf{x})) = 0$. If $\mathbf{x} \in V$ then $f(\mathbf{x}) > 0$ so $f_{(K,U)}(\mathbf{x}) = h_{(0,r)}(f(\mathbf{x})) > 0$. Hence, $spt(f_{(K,U)}) = \overline{V} \subset U$.

□

To be more specific, $spt(f_{(K,U)}) = \displaystyle\bigcup_{i=1}^{m} R_{2\epsilon}(\mathbf{x}_i)$ in the previous argument, but equality was not needed for the result in question.

It is helpful for certain theorems to have the notion of and F_σ and G_δ set.

Definition 130

The intersection of countably many open sets is a G_δ set. The union of countably many closed sets is an F_σ set. If a set S is the union of countably many compact sets then the set is referred to as being σ-compact.

Now, in $\mathbb{R}^n$ there isn't any difference between the set of sets that are F_σ and the set of sets that are σ-compact.

Theorem 14.19. *Every closed set is a G_δ set in $\mathbb{R}^n$. For each open set U there are compact sets $\{K_i\}_{i\in\mathbb{N}}$ so that $K_1 \subset K_2^\circ \subset K_2 \subset K_3^\circ \subset K_3 \subset K_4^\circ ...$ so that $U = \bigcup_{i=1}^{\infty} K_i$.*

Proof. First, let A be a closed set. Let $V_i = \bigcup_{\mathbf{x}\in A} B_{\frac{1}{i}}(\mathbf{x})$. Note that $V_i \subseteq V_j$ if $i > j$ and $A \subseteq \bigcap_{i=1}^{\infty} V_i$ since every point of A is contained in an open ball contained in V_i for each $i \in \mathbb{N}$. Let $\mathbf{p} \notin A$. Then since A is closed we know that $B_{\frac{1}{m}}(\mathbf{p}) \cap A = \emptyset$. Hence, there is no point of A within distance $\dfrac{1}{m}$ of $\mathbf{p}$ and therefore $\mathbf{p} \notin V_m$. Thus, $A = \bigcap_{i=1}^{\infty} V_i$, so every closed set is a G_δ set.

Let U be an open set. Let $V_i = \bigcup_{\mathbf{x}\in\mathbb{R}^n\setminus U} B_{\frac{1}{i}}(\mathbf{x})$. By the preceding argument we know that $\bigcap_{i=1}^{\infty} V_i = \mathbb{R}^n \setminus U$, so by Demorgan's Laws we know that $\bigcup_{i=1}^{\infty} \mathbb{R}^n \setminus V_i = U$. Let $K_m = \overline{B_m(\mathbf{0})} \cap \bigcup_{i=1}^{m} \mathbb{R}^n \setminus V_i = \overline{B_m(\mathbf{0})} \cap \mathbb{R}^n \setminus V_m$ for each $m \in \mathbb{N}$. Since each K_m is an intersection of closed sets and is therefore closed (and bounded since $K_m \subseteq \overline{B_m(\mathbf{0})}$), it follows from the Heine-Borel Theorem that K_m is compact.

For every $\mathbf{x} \in U$ there is some k so that $\mathbf{x} \in B_k(\mathbf{0})$ and $\mathbf{x} \in \mathbb{R}^n \setminus V_k$, so $\bigcup_{n=1}^{\infty} K_i = U$. Let $\mathbf{p} \in K_m$ for some m. Let $0 < \delta < \dfrac{1}{m} - \dfrac{1}{m+1}$ and let $\mathbf{x} \in B_\delta(\mathbf{p})$. Then $|\mathbf{x} - \mathbf{0}| \leq |\mathbf{x} - \mathbf{p}| + |\mathbf{p} - \mathbf{0}| < 1 + m$, which means that $\mathbf{x} \in \overline{B_{m+1}(\mathbf{0})}$. Also, if $\mathbf{y} \in \mathbb{R}^n \setminus U$ then $|\mathbf{y} - \mathbf{x}| \geq |\mathbf{y} - \mathbf{p}| - |\mathbf{p} - \mathbf{x}| \geq \dfrac{1}{m} - \delta > \dfrac{1}{m+1}$, which means that $\mathbf{x} \in \mathbb{R}^n \setminus V_{m+1}$. Hence, $B_\delta(\mathbf{p}) \subseteq K_{m+1}$ and $\mathbf{p} \in K_{m+1}^\circ$. It follows that $K_1 \subset K_2^\circ \subset K_2 \subset K_3^\circ \subset K_3 \subset K_4^\circ ....$ □

An immediate consequence of this theorem is that open sets are σ-compact (and thus F_σ sets).

We now define a useful idea for breaking a function down to a sum of simpler functions called a partition of unity. This will help us with the change of variables theorem. In general, it is useful for purposes of proofs extending local properties of functions to global properties.

Definition 131

Let $S \subseteq \mathbb{R}^n$. Let U be an open set containing S. Then a set Φ of C^∞ functions $\phi_\beta : U \to \mathbb{R}$, for β in some indexing set W, is a C^∞ *partition of unity* for S if Φ satisfies the following conditions:
(1) If $\mathbf{x} \in S$ and $\phi_\beta \in \Phi$ then $\phi_\beta(\mathbf{x}) \in [0, 1]$.
(2) For each $\mathbf{p} \in S$ there is an $\epsilon_\mathbf{p} > 0$ so that $F_{(\Phi,\epsilon_\mathbf{p})} = \{\phi_\beta \in \Phi | spt(\phi_\beta) \cap B_{\epsilon_\mathbf{p}}(\mathbf{p}) \neq \emptyset\}$ is finite.
(3) For each $\mathbf{x} \in S$, the sum $\displaystyle\sum_{\beta \in W} \phi_\beta(\mathbf{x}) = \sum_{\phi_\beta \in F_{(\Phi,\epsilon_\mathbf{p})}} \phi_\beta(\mathbf{x}) = 1$.
If we replace ∞ by a positive integer p in the preceding definition then Φ is called a C^p partition of unity for S.
Let $C = \{U_\alpha\}_{\alpha \in J}$ be an open cover of S. If Φ also satisfies the following condition then Φ is a partition of unity *subordinate to* C.
(4) For each $\phi_\beta \in \Phi$ there is some $U_\alpha \in C$ so that $spt(\phi_\beta) \subset U_\alpha$.

A partition of unity for a set subordinate to a cover of that is a also a partition of unity for one of its subsets subordinate to the same cover. We go through this observation below.

Theorem 14.20. *Let $S \subseteq \mathbb{R}^n$, let $R \subseteq S$ and let $\Phi = \{\phi_\beta\}_{\alpha \in \gamma}$ be a partition of unity for S subordinate to the cover $C = \{U_\alpha\}_{\alpha \in J}$, where each $\phi_\beta : U \to \mathbb{R}$ and U is open in $\mathbb{R}^n$. Let D be an open cover of R so that each element of C is contained in an element of D. Then Φ is a partition of unity for R subordinate to D.*

Proof. We check each condition.

(1) If $\mathbf{x} \in R$ then $\mathbf{x} \in S$, so for every $\phi_\beta \in \Phi$ we know $\phi_\beta(\mathbf{x}) \in [0, 1]$.

(2) If $\mathbf{p} \in R$ then $\mathbf{p} \in S$, so there is an $\epsilon_\mathbf{p} > 0$ so that $F_{(\Phi,\epsilon_\mathbf{p})} = \{\phi_\beta \in \Phi | spt(\phi_\beta) \cap B_{\epsilon_\mathbf{p}}(\mathbf{p}) \neq \emptyset\}$ is finite.

(3) For each $\mathbf{x} \in R$, we know $\mathbf{x} \in S$, so the sum $\displaystyle\sum_{\beta \in W} \phi_\beta(\mathbf{x}) = \sum_{\phi_\beta \in F_{(\Phi,\epsilon_\mathbf{p})}} \phi_\beta(\mathbf{x}) = 1$.

(4) Since Φ is subordinate to C we know that for each $\phi_\beta \in \Phi$ there is some $U_\alpha \in C$ so that $spt(\phi_\beta) \subset U_\alpha \subseteq V$ for some $V \in D$.

Hence, we know that Φ is a partition of unity for R subordinate to D. □

We next prove that we can find a partition of unity subordinate to any cover of any set.

Theorem 14.21. *Let $C = \{U_\alpha\}_{\alpha \in J}$ be an open cover of $S \subseteq \mathbb{R}^n$. Then there is a countable (finite if S is compact) C^∞ partition of unity $\Phi = \{\phi_\beta\}_{\beta \in D}$ for S which is subordinate to C.*

Proof. We first show that we can find a finite partition of unity subordinate to C assuming that S is compact.

By the Lebesgue Number Lemma we can find $\epsilon > 0$ so that if a set Q of diameter less than or equal to 4ϵ intersects S then $Q \subset U_\alpha$ for some $U_\alpha \in C$ and also $Q \subset U$. Let $D = \{B_\epsilon(\mathbf{x})\}_{\mathbf{x}\in S}$. Since S is compact, D has a finite subset $E = \{B_\epsilon(\mathbf{x}_i)\}_{1\leq i \leq m}$ which covers S. By Theorem 14.18, for each $i \in \{1,2,3,...,m\}$ we can find an open set V_i and a C^∞ function $f_{(\overline{B_\epsilon(\mathbf{x}_i)},B_{2\epsilon}(\mathbf{x}_i))} : \mathbb{R}^n \to [0,1]$ so that $\overline{B_\epsilon(\mathbf{x}_i)} \subset V_i \subset \overline{V_i} = spt(f_{(\overline{B_\epsilon(\mathbf{x}_i)},B_{2\epsilon}(\mathbf{x}_i))}) \subset B_{2\epsilon}(\mathbf{x}_i)$, where $f_{(\overline{B_\epsilon(\mathbf{x}_i)},B_{2\epsilon}(\mathbf{x}_i))}(\mathbf{x}) = 1$ if $\mathbf{x} \in \overline{B_\epsilon(\mathbf{x}_i)}$, $f_{(\overline{B_\epsilon(\mathbf{x}_i)},B_{2\epsilon}(\mathbf{x}_i))}(\mathbf{x}) > 0$ if and only if $\mathbf{x} \in V_i$ and $B_{2\epsilon}(\mathbf{x}_i) \subset U_{\alpha_i}$ for some $\alpha_i \in J$. Setting $V = \bigcup_{i=1}^m V_i$, it follows that $S \subset V \subset \overline{V} \subset U$.

Let $W = \bigcup_{i=1}^m B_\epsilon(\mathbf{x}_i)$. By Theorem 14.18, we can find a C^∞ function $f_{(S,W)} : \mathbb{R}^n \to [0,1]$ so that $f_{(S,W)}(\mathbf{x}) = 1$ if $\mathbf{x} \in S$ and $spt(f_{(S,W)}) \subset W$.

For each natural number $i \leq m$, we define a function $\phi_i : \mathbb{R}^n \to [0,1]$ by $\phi_i(\mathbf{x}) = \dfrac{f_{(S,W)}(\mathbf{x})f_{(\overline{B_\epsilon(\mathbf{x}_i)},B_{2\epsilon}(\mathbf{x}_i))}(\mathbf{x})}{\sum_{j=1}^m f_{(\overline{B_\epsilon(\mathbf{x}_j)},B_{2\epsilon}(\mathbf{x}_j))}(\mathbf{x})}$. The denominator $\sum_{j=1}^m f_{(\overline{B_\epsilon(\mathbf{x}_j)},B_{2\epsilon}(\mathbf{x}_j))}(\mathbf{x})$ is non-zero on V and $W \subset V$, so each ϕ_i is C^∞ on W. If $\mathbf{x} \notin W$ then $\mathbf{x} \notin spt(f_{(S,W)})$, so we can find some $\delta > 0$ so that $B_\delta(\mathbf{x}) \cap spt(f_{(S,W)}) = \emptyset$ (since $spt(f_{(S,W)})$ is closed). Since $f_{(S,W)}$ is zero on $B_\delta(\mathbf{x})$ we know that ϕ_i is also zero on $B_\delta(\mathbf{x})$, which means that ϕ_i is C^∞ at $\mathbf{x}$. Hence ϕ_i is C^∞ on $\mathbb{R}^n$ as desired.

If $\mathbf{x} \in S$ then $f_{(S,W)}(\mathbf{x}) = 1$, so $\displaystyle\sum_{i=1}^m \phi_i(\mathbf{x}) = \frac{\sum_{j=1}^m f_{(\overline{B_\epsilon(\mathbf{x}_j)},B_{2\epsilon}(\mathbf{x}_j))}(\mathbf{x})}{\sum_{j=1}^m f_{(\overline{B_\epsilon(\mathbf{x}_j)},B_{2\epsilon}(\mathbf{x}_j))}(\mathbf{x})} = 1$.

Next, we extend this to an arbitrary set S. Let $D = \bigcup C$. By Theorem 14.19, we can find a sequence of compact sets $K_1 \subset K_2^\circ \subset K_2 \subset K_3^\circ ...$ so that $\bigcup_{i=1}^\infty K_i = D$. By the argument above, we can find finite partitions of unity Φ_1 for K_1 and Φ_2 for K_2 subordinate to C, and Φ_i for each natural number $i \geq 3$, so that Φ_i is a partition of unity for $K_i \setminus K_{i-1}^\circ$ which is subordinate to the open cover $C_i = \{U_\alpha \cap (K_{i+1}^\circ \setminus K_{i-2}) | U_\alpha \in C\}$ of $K_i \setminus K_{i-1}^\circ$, where each $\phi \in \bigcup_{i=1}^\infty \Phi_i$ is a C^∞ function so that $\phi : \mathbb{R}^n \to [0,1]$. Let $\Phi = \bigcup_{i=1}^\infty \Phi_i$.

Since Φ is countable, we can list $\Phi = \{\phi_1, \phi_2, \phi_3, ...\}$. If $\mathbf{x} \in D$ then let j be the first natural number so that $\mathbf{x} \in K_j^\circ$. Then $\phi_r(\mathbf{x}) > 0$ for some $\phi_r \in \Phi_j$. Choose $\gamma > 0$ so that $B_\gamma(\mathbf{x}) \subseteq K_j^\circ$. For all $i \geq j+2$ we know that $B_\gamma(\mathbf{x}) \cap (K_{i+1}^\circ \setminus K_{i-2}) = \emptyset$ which means that there are only finitely many ϕ_i which are non-zero on $B_\gamma(\mathbf{x})$. Hence $s(\mathbf{x}) = \sum_{i=1}^\infty \phi_i(\mathbf{x})$ is a finite sum of C^∞ functions on $B_\delta(\mathbf{x})$, which means that $s(\mathbf{x})$ is a C^∞ positive function on D. For each $i \in \mathbb{N}$, define $\psi_i : D \to \mathbb{R}$ by $\psi_i(\mathbf{x}) = \dfrac{\phi_i(\mathbf{x})}{s(\mathbf{x})}$, and let $\Psi = \{\psi_i\}_{i\in\mathbb{N}}$. Since s is positive on D it follows that each ψ_i is C^∞, $\dfrac{\sum_{i=1}^\infty \psi(\mathbf{x})}{s(\mathbf{x})} = 1$ for all $\mathbf{x} \in D$, and $spt(\psi_i) = spt(\phi_i)$ which is a subset of an element of C for each $i \in \mathbb{N}$, so Ψ is a partition of unity for D subordinate to C, and therefore a partition of unity for S subordinate to C by Theorem 14.20.

□

Theorem 14.22. *Let $C = \{U_\alpha\}_{\alpha \in J}$ be an open cover of $S \subseteq \mathbb{R}^n$ with C^∞ partition of unity $\Phi = \{\phi_1, \phi_2, \phi_3, ...\}$ for S where each ϕ_i is defined on some open set $U \supseteq S$ and Φ is subordinate to C. Let K be a compact subset of S. Then $\Phi_K = \{\phi_i \in \Phi | \phi(\boldsymbol{x}) > 0$ for some $\boldsymbol{x} \in K\}$ is finite.*

Proof. For each $\mathbf{p} \in K$ choose $\epsilon_\mathbf{p} > 0$ so that $F_{(\Phi, \epsilon_\mathbf{p})} = \{\phi_i \in \Phi | spt(\phi_i) \cap B_{\epsilon_\mathbf{p}}(\mathbf{p}) \neq \emptyset\}$ is finite. Then $\{B_{\epsilon_\mathbf{p}}(\mathbf{p})\}_{\mathbf{p} \in K}$ is an open cover of K which has a finite subcover $F = \{B_{\epsilon_{\mathbf{p}_i}}(\mathbf{p})\}_{1 \le i \le m}$. Since there are only finitely many ϕ_i which are non-zero on each $B_{\epsilon_{\mathbf{p}_j}}(\mathbf{p}_j)$, it follows that there are only finitely many ϕ_i which are non-zero on K. □

We mentioned that partitions of unity are useful for taking local conditions and piecing them together to form global conditions. We conclude with an example to illustrate this notion.

Definition 132

We say an open cover $\mathcal{C}$ of an open set U in $\mathbb{R}^n$ is *admissible* if $W \subset U$ for every $W \in \mathcal{C}$. Let Φ by a partition of unity subordinate to $\mathcal{C}$. Let $f : U \to \mathbb{R}$ be a function so that for each $\mathbf{x} \in U$ there is some $\epsilon_\mathbf{x} > 0$ so that f is bounded on $B_{\epsilon_\mathbf{x}}(\mathbf{x})$ and the set of discontinuities of f has measure zero. We define f to be *integrable in the extended sense* if $\displaystyle\sum_{\phi \in \Phi} \int_U \phi|f|$ converges. In this case, we define $\displaystyle\int_U f = \sum_{\phi \in \Phi} \int_U \phi f$.

Note that there is no problem with using $\displaystyle\int_U \phi|f|$ in the definition above since each ϕ is zero outside of a Jordan region, and this notation just means the integral over a Jordan region contained in U on which f is non-zero. Also, the absolute convergence of $\displaystyle\sum_{\phi \in \Phi} \int_U \phi f$ is implied by the convergence of $\displaystyle\sum_{\phi \in \Phi} \int_U \phi|f|$. One can demonstrate that this definition is independent of the particular partition of unity Φ and cover $\mathcal{C}$ chosen, but we will not be pursuing this topic further in this text.

The above definition allows us to extend the notion of integrability to an arbitrary open set rather than just a Jordan region, which is often interesting. A simple example of an instance where such an idea is of interest for single variable functions is the idea of an improper integral.

REFERENCES

1. Michael Spivak, *Calculus on Manifolds*, Westview Press, 1965.

2. William R. Wade, *An Introduction to Analysis*, fourth edition, Prentice Hall, 2009

3. R. Creighton Buck, *Advanced Calculus*, International Series in Pure and Applied Mathematics, 1956

4. Richard Courant, Fritz John, *Introduction to Calculus and Analysis, Volume II*, Springer-Verlag, 1974.

Index